

CONVEXITY AND GEOMETRY OF ESTIMATING FUNCTIONS

By

SCHULTZ CHAN

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1996

TABLE OF CONTENTS

ABSTRACT	iv
1 INTRODUCTION	1
1.1 Preamble	1
1.2 Literature Review	1
1.3 The Subject of the Dissertation	4
2 THE GEOMETRY OF ESTIMATING FUNCTIONS I	6
2.1 Introduction	6
2.2 Generalized Inner Product Spaces and Orthogonal Projection	7
2.3 Optimal Estimating Functions: A Geometric Approach	16
2.4 Optimal Bayesian Estimating Functions	24
2.5 Orthogonal Decomposition and Information Inequality	32
2.6 Orthogonal Decomposition for Estimating Functions	35
3 THE GEOMETRY OF ESTIMATING FUNCTIONS II	40
3.1 Introduction	40
3.2 Properties of Orthogonal Projections	41
3.3 Global Optimality of Estimating Functions	42
3.3.1 The General Result	43
3.3.2 Geometry of Conditional Inferences	45
3.3.3 Geometry of Marginal Inference	48
3.4 Locally Optimal Estimating Functions	51
3.4.1 A General Result	52
3.4.2 Local Optimality of Conditional Score Functions	53
3.4.3 Locally Optimal Estimating Functions for Stochastic Processes	55
3.4.4 Local Optimality of Projected Partial Likelihood	58
3.5 Optimal Conditional Estimating Functions	61
4 CONVEXITY AND ITS APPLICATIONS TO STATISTICS	66
4.1 Introduction	66

4.2	Some Simple Results About Convexity	67
4.3	Theory of Optimum Experimental Designs	72
4.4	Fundamental Theorem of Mixture Distributions	77
4.5	Asymptotic Minimality of Estimating Functions	79
4.5.1	One Dimensional Case	82
4.5.2	Multi-Dimensional Case	85
5	SUMMARY AND FUTURE RESEARCH	89
5.1	Summary	89
5.2	Future Research	89
	BIBLIOGRAPHY	91
	BIOGRAPHICAL SKETCH	97

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy

CONVEXITY AND GEOMETRY OF ESTIMATING FUNCTIONS

By

Schultz Chan

August 1996

Chairman: Malay Ghosh
Major Department: Statistics

In this dissertation, a general way of constructing optimal generalized estimating equations (GEE) is given. Applications of this general method to various statistical problems, such as the quasi-likelihood method in generalized linear models, Cox's partial likelihood method in survival analysis, Bayesian inference, conditional and marginal inferences, are also studied. Also, some simple results about matrix valued convex functions are proved and are applied to the study of optimal designs, mixture distributions and asymptotic minimaxity.

First, a notion of generalized inner product spaces is introduced to study optimal estimating functions. A characterization of orthogonal projections in generalized inner product spaces is given. It is shown that the orthogonal projection of the score function into a linear subspace of estimating functions is optimal in that subspace, and a characterization of optimal estimating functions is given. Also optimal estimating functions in the Bayesian framework are also studied.

In the case of no nuisance parameters, the results are applied to study longitudinal data, stochastic processes, time series models, generalized linear models and Bayesian inference. As special cases of the main results of this chapter, we derive the results of Godambe on the foundation of estimation in stochastic processes, the result of Godambe and Thompson on the extension of quasi-likelihood, and the linear (and quadratic) generalized estimating equations for multivariate data due to Liang and Zeger, Liang, Zeger and Qaqish. Also we have derived optimal Bayesian estimating equations in the Bayesian framework.

In the case where there are nuisance parameters, the results are applied to study survival analysis models, the generalized estimating equations proposed by Liang, Zeger and their associates, and the optimality of the marginal and conditional inferences. The three main topics are (A) globally optimal generalized estimating equations; (B) locally optimal generalized estimating equations; (C) conditionally optimal generalized estimating equations. A general result is derived in each case. As special cases, we rederive some of the results already available in the literature and find also some new results. In particular, as special cases of our result on globally optimal generalized estimating equations, we find the results of Godambe and Thompson and Godambe with nuisance parameters. The results of Bhapkar on conditional and marginal inference are also obtained as special cases. As applications of our result on locally optimal generalized estimating equations, we find Lindsay's result on the optimality of conditional score functions, extend Godambe's result on optimal estimating functions for stochastic processes to nuisance parameters, and extend a recent result of Murphy and Li about projected partial likelihood. Finally, our general result on conditionally optimal generalized estimating equation helps generalize the findings of Godambe and Thompson to situations which admit the presence of nuisance parameters.

Finally, some simple results for matrix valued convex functions are proved, and are used to find optimum experimental designs, the fundamental theorem of mixture distributions, and a generalization of the asymptotic result of Huber.

CHAPTER 1

INTRODUCTION

1.1 Preamble

The objective of this thesis is to provide a geometric insight behind many useful concepts in statistics and utilize the geometry for unifying many existing results, as well as in deriving several new ones. One major focus is to find optimal estimating functions as orthogonal projections of score functions into appropriate linear subspaces. The second goal is to use some important theorems from convex analysis for finding optimal experimental designs, for deriving the fundamental theorem of mixture distributions, and for proving the asymptotic minimaxity of estimating functions in a very general framework.

1.2 Literature Review

We begin by reviewing the literature on estimating functions. The topic has grown into an active research area over the past decade. Its beginning is marked with the celebrated articles of Godambe (1960) and Durbin (1960). While Durbin (1960) used estimating functions to study Gauss-Markov type results in a time series setting, Godambe's (1960) main objective was to prove the optimality of the score function in a parametric framework when there were no nuisance parameters. As is well-known, the Gauss-Markov theory and maximum likelihood estimation form two cornerstones of statistical estimation. In their review article, Godambe and Kale

(1991) have pointed out that the theory of estimating functions combines the strengths of these two methods, eliminating at the same time many of their weaknesses. To cite an example, Gauss-Markov theorem fails for nonlinear least squares, but estimators obtained as solutions of optimal estimating equations are identical to the least squares estimators under homoscedasticity.

The theory of estimating functions has made rapid strides since the 1970s. Godambe and Thompson (1974), Godambe (1976) studied optimal estimating functions in the presence of nuisance parameters, and proved a variety of optimality results. Bhapkar (1972, 1989, 1991), Bhapkar and Srinivasan (1994) in a series of articles studied the notions of sufficiency, ancillarity and information in the context of estimating functions, and found conditional as well as marginal optimal estimating functions. Amari and Kumon (1988), and Kumon and Amari (1984) used estimating functions to estimate structural parameters in the presence of a large number of nuisance parameters, their approach being based on vector bundle theory from differential geometry.

Nelder and Wedderburn (1971), in their pioneering paper on generalized linear models, showed that using one algorithm (the Newton-Raphson method), a large family of models could be iteratively fitted. Later, Wedderburn (1974) realized that only the first two moments were utilized in fitting the models, and this led to the development of the so-called quasi-likelihood functions for the development of generalized linear models. Firth (1987), and Godambe and Thompson (1989) pointed out the connection between quasi-likelihood and optimal estimating functions. An interesting review article is due to Desmond (1991).

Cox (1972), in his seminal paper, introduced the proportional hazards model. Later, Cox (1975) introduced the notion of *partial likelihood*. The latter is intended to eliminate nuisance parameters (baseline hazards for the proportional hazards model)

by using a conditioning argument. Because of the nested structure of the conditioning variables, Cox's approach also fits into the estimating function framework.

Liang and Zeger (1986) used estimating functions (they used the terminology generalized estimating equations) to study longitudinal data. Liang and Zeger had motivation similar to Wedderburn's quasi-likelihood function, but in the multivariate setting in order to take into account the correlation between responses within each subject.

Bayesian estimating function is of more recent origin and is still in its infancy. Ferreira (1981, 1982) and Ghosh (1990) initiated the study of optimal estimating functions in a Bayesian framework. While Ferreira's formulation involves the joint distribution of the observations and the parameters, Ghosh used a pure Bayesian approach based only on the posterior probability density function.

The theory of optimum experimental designs, it was initiated by Elfving (1952, 1959), and Kiefer (1959). For the references up to the early eighties, we refer to the two monographs of Silvey (1980) and Pazman (1986). During the last decade, there are major advances in optimum experimental design theory, here we only list a few of the main publications. Chaloner and Larntz (1989) studied optimal Bayesian design for logistic regression model, El-Krunz and Studden (1991) studied Bayesian optimal design for linear regression models, while DasGupta and Studden (1991) studied robust Bayesian experimental designs for normal linear models. Dette and Studden (1993) studied the geometry of E -optimal design, while Dette (1993) studied the geometry of D -optimal design, and Haines (1995) studied the geometry of Bayesian designs.

There is a vast literature on the theory of mixture distributions. Laird (1978) studied nonparametric maximum likelihood estimation of a mixing distribution. Lindsay (1981, 1983a, 1983b) studied the properties and geometry of maximum likelihood estimator of mixing distribution. In a recent monograph, Lindsay (1995) presented

a comprehensive treatment of “mixture models: theory, geometry and applications.” In this book, a variety of topics about mixture distributions were discussed, which include the well known result proved by Shaked (1980) on mixtures from the exponential family, and the fundamental theorem on mixture distributions proved by Lindsay (1983a).

Huber (1964) in his pioneering paper, proved the well known asymptotic minimaxity result for estimating functions about location parameter. In his classical book on robust statistics, Huber (1980) presented a more systematic treatment about asymptotic minimaxity.

1.3 The Subject of the Dissertation

This dissertation begins with unfolding the geometry of estimating functions, and pointing out many applications. Although the geometry is primarily used to study estimating functions, this can also be used to study other statistical topics, such as the Rao-Blackwell theorem, Lehmann-Scheffe’s approach to uniform minimum variance unbiased estimators and predication theory.

In Chapter 2, a notion of generalized inner product spaces is introduced to study optimal estimating functions. A characterization of orthogonal projections in generalized inner product spaces is given. It is shown that the orthogonal projection of the score function into a linear subspace of estimating functions is optimal in that subspace and a characterization of optimal estimating functions are given. As special cases of the main results of this paper, we derive the results of Godambe (1985) on the foundation of estimation in stochastic processes, the result of Godambe and Thompson (1989) on the extension of quasi-likelihood, and the generalized estimating equations for multivariate data due to Liang and Zeger (1986). Also we have derived optimal estimating functions in the Bayesian framework. This generalizes the results obtained by Ferreira (1981, 1982) and Ghosh (1990).

In Chapter 3, the geometry of estimating functions in the presence of nuisance parameters is studied. The three main topics are: (A) globally optimal estimating functions; (B) locally optimal estimating functions; (C) conditionally optimal estimating functions. A general result is derived in each case. As special cases, we rederive some of the results already available in the literature, and find also some new results. In particular, as special cases of our result on globally optimal estimating functions, we find the results of Godambe and Thompson (1974) and Godambe (1976) with nuisance parameters. The results of Bhapkar (1989, 1991a) on conditional and marginal inference are also obtained as special cases. As applications of our result on locally optimal estimating functions, we find Lindsay's (1982) result on the optimality of conditional score functions, extend Godambe's (1985) result on optimal estimating functions for stochastic processes, and extend a recent result of Murphy and Li (1995) about projected partial likelihood. Finally, our general result on conditionally optimal estimating function helps generalize the findings of Godambe and Thompson (1989) to situations which admit the presence of nuisance parameters.

In Chapter 4, we first prove some general results about convexity, and then apply the results to various statistical problems, which include the theory of optimum experimental designs (Silvey, 1980), the fundamental theorem of mixture distributions due to Lindsay (1983a), and the asymptotic minimaxity of robust estimation due to Huber (1964). In his classical paper on M -estimation, Huber (1964) proved an asymptotic minimaxity result for estimating functions about a location parameter. In this chapter, this fundamental result is generalized to general estimating functions. The geometric optimality of estimating functions proved in Chapter 2 will be used to prove a necessary and sufficient condition for the asymptotic minimaxity of estimating functions when the parameter space is multi-dimensional.

In Chapter 5, we summarize the results of this dissertation, and propose some topics of future research.

CHAPTER 2

THE GEOMETRY OF ESTIMATING FUNCTIONS I

2.1 Introduction

The theory of estimating functions has advanced quite rapidly over the past two decades. Godambe (1960) introduced the subject to prove finite sample optimality of the score function in a parametric framework when no nuisance parameters were presented. Later, his idea was extended in many different directions, and optimal estimating functions were derived under many different formulations.

The underlying thread in all these results is a geometric phenomenon which seems to have gone unnoticed, or at least has never been brought out explicitly. In the present chapter, we make this geometry explicit, and use the same in deriving optimal estimating functions in certain contexts. In particular, it is shown that optimal estimating functions for certain semiparametric models are indeed the orthogonal projections of score functions into certain linear subspaces. Also, this geometry, by its very nature, is neutral, and can be adapted both within the frequentist, and the Bayesian paradigm. Second, the multiparameter situation can be handled automatically through this geometry without involving any additional work.

The outline of the remaining sections is as follows. In Section 2.2, we develop the mathematical prerequisite for the results of the subsequent sections. In particular, we define generalized inner product spaces, and show the existence of orthogonal

projections of elements in these spaces into some linear subspaces. A characterization theorem for these orthogonal projections is given, which is used repeatedly in subsequent sections.

Section 2.3 generalizes the results of Godambe (1985) and Godambe and Thompson (1989) in multiparameter situations and also finds optimal generalized estimating equations (GEEs) for multivariate data. The GEEs used in Liang and Zeger (1986) and Liang, Zeger and Qaqish (1992) turn out to be special cases of those proposed in this section. The common thread in the derivation of all the optimal estimating functions is the idea of orthogonal projection developed in section 2.2.

Section 2.4 uses the orthogonal projection idea in deriving optimal Bayes estimating functions. The results of Ferreira (1981, 1982) and Ghosh (1990) are included as special cases. Section 2.5 uses an orthogonal decomposition to study information inequality. Section 2.6 studies the Hoeffding type decomposition for estimating functions.

2.2 Generalized Inner Product Spaces and Orthogonal Projection

In this section, we first introduce a matrix version of inner product spaces which generalizes the notion of the usual scalar valued inner product space. Next we provide the definition of the orthogonal projection of an element of a generalized inner product space, say, L into a linear subspace L_0 of L . A characterization of the orthogonal projection in the generalized inner product space is also given. As will be seen, such a characterization generalizes a corresponding result for scalar inner product spaces. We also show that for a finite dimensional subspace of a generalized inner product space, an orthogonal projection always exists.

We begin with the definition of a matrix valued inner product space.

Definition 2.2.1. Let L be a real linear space, and let $M_{k \times k}$ be the set of all $k \times k$ real matrices. The map

$$\langle \cdot, \cdot \rangle : L \times L \longrightarrow M_{k \times k},$$

is called a generalized inner product if

$$(1) \forall x, y \in L, \langle x, y \rangle = \langle y, x \rangle^t;$$

$$(2) \text{ for any } k \times k \text{ matrix } M, x, y \in L, \langle Mx, y \rangle = M \langle x, y \rangle;$$

$$(3) \forall x, y, z \in L, \langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle;$$

$$(4) \forall x \in L, \langle x, x \rangle \text{ is non negative definite (n. n. d.), and } \langle x, x \rangle = 0 \text{ iff } x = 0.$$

Two elements $x, y \in L$ are said to be orthogonal if $\langle x, y \rangle = 0$. Two sets S_1, S_2 are orthogonal if every element of S_1 is orthogonal to every element of S_2 .

An example of a generalized inner product space, of great interest to statisticians is the one where the generalized inner product is defined by the covariance matrix of random vectors. Specifically, let \mathcal{X} be a sample space, and let $\Theta \subset R^k$ be the parameter space, which is open. Consider the space L of all functions

$$h : \mathcal{X} \times \Theta \longrightarrow R^k,$$

such that every element of the matrix $E[h(X, \theta)h(X, \theta)^t | \theta]$ is finite. For any $h, g \in L, \theta \in \Theta$, the family of generalized inner products is defined by

$$\langle h, g \rangle_\theta = E[h(X, \theta) g(X, \theta)^t | \theta].$$

Then it is easy to verify that for fixed $\theta \in \Theta, \langle \cdot, \cdot \rangle_\theta$ is a generalized inner product on L .

Definition 2.2.2. Let L be a generalized inner product space with inner product $\langle \cdot, \cdot \rangle$. Suppose L_0 is a linear subspace of L . Let $s \in L$. An element $y_0 \in L_0$ is called the orthogonal projection of s into L_0 if

$$\langle s - y_0, s - y_0 \rangle = \min_{y \in L_0} \langle s - y, s - y \rangle, \quad (2.2.1)$$

where \min is taken with respect to the usual ordering of matrices. More specifically, for two square matrices A and B of the same order, we say that $A \geq B$ if $A - B$ is n. n. d.

The following theorem characterizes the orthogonal projection in generalized inner product spaces.

Theorem 2.2.1. Let L be a generalized inner product space with inner product $\langle \cdot, \cdot \rangle$, and L_0 be a linear subspace of L . Let $s \in L$. Then $y_0 \in L_0$ is the orthogonal projection of s into L_0 if and only if

$$\langle s - y_0, y \rangle = 0, \quad (2.2.2)$$

for all $y \in L_0$, i. e., $s - y_0$ and L_0 are orthogonal. Furthermore, if the orthogonal projection exists, then it is unique.

Proof. Only if. For all $y \in L_0, \alpha \in R$, since $y_0 - \alpha y \in L_0$,

$$\langle s - y_0 + \alpha y, s - y_0 + \alpha y \rangle - \langle s - y_0, s - y_0 \rangle$$

is n. n. d., i. e.,

$$\alpha[\langle s - y_0, y \rangle + \langle s - y_0, y \rangle^t] + \alpha^2 \langle y, y \rangle \quad (2.2.3)$$

is n. n. d., for all $y \in L_0, \alpha \in R$.

Now suppose that there exists $y_0^* \in L_0$ such that $\langle s - y_0, y_0^* \rangle \neq 0$. Let

$$A = \langle s - y_0, y_0^* \rangle + \langle s - y_0, y_0^* \rangle^t.$$

Then A is real symmetric, and $A \neq 0$. Suppose $\lambda_1, \dots, \lambda_k$ are the eigenvalues of A with $|\lambda_1| \geq \dots \geq |\lambda_k|$. Denote by z_1 the unit eigenvector corresponding to λ_1 . Then from (3), using $z_1^t A z_1 = \lambda_1$,

$$\alpha \lambda_1 + \alpha^2 z_1^t \langle y_0, y_0 \rangle z_1 \geq 0,$$

for all $\alpha \in R$. This implies that $\lambda_1 = 0$. So $A = 0$, a contradiction. Hence,

$$\langle s - y_0, y \rangle = 0,$$

for all $y \in L_0$.

If. Suppose

$$\langle s - y_0, y \rangle = 0,$$

for all $y \in L_0$. Then

$$\begin{aligned} \langle s - y, s - y \rangle &= \langle s - y_0, s - y_0 \rangle \\ &= \langle s - y_0 + y_0 - y, s - y_0 + y_0 - y \rangle = \langle s - y_0, s - y_0 \rangle \\ &\quad + \langle y_0 - y, y_0 - y \rangle, \end{aligned} \tag{2.2.4}$$

which is n. n. d. The last equality follows since $\langle s - y_0, y_0 - y \rangle = 0$. This implies

$$\langle s - y_0, s - y_0 \rangle = \min_{y \in L_0} \langle s - y, s - y \rangle.$$

Finally we show that if an orthogonal projection exists, then it is unique. Suppose that $y_1, y_2 \in L_0$ are both orthogonal projections of s into L_0 . Then

$$\langle s - y_i, y \rangle = 0,$$

for all $y \in L_0$, $i = 1, 2$. In particular,

$$\begin{aligned} \langle y_1 - y_2, y_1 - y_2 \rangle &= \langle s - y_2, y_1 - y_2 \rangle - \langle s - y_1, y_1 - y_2 \rangle \\ &= 0 - 0 = 0. \end{aligned}$$

So $y_1 = y_2$. This completes the proof of the theorem.

Next we apply Theorem 2.2.1 to generalize a result of Lehmann-Scheffe to the multidimensional case. Let \mathcal{X} be a sample space, $\Theta \subset R^k$ an open set, and $\gamma : \Theta \rightarrow R^d$ an estimable function, i. e., there exists $g : \mathcal{X} \rightarrow R^d$ such that

$$E[g(X)|\theta] = \gamma(\theta), \quad \forall \theta \in \Theta.$$

Let

$$U_\gamma = \{g : \mathcal{X} \rightarrow R^d | E[g(X)|\theta] = \gamma(\theta), \quad \forall \theta \in \Theta\},$$

$$U_0 = \{h : \mathcal{X} \longrightarrow R^d | E[h(X)|\theta] = 0, \quad \forall \theta \in \Theta\},$$

where $g \in U_\gamma, h \in U_0$ satisfy that $E[g(X) g(X)^t | \theta]$ and $E[h(X) h(X)^t | \theta]$ are all well defined. Note that $g_* \in U_\gamma$ is a locally minimum variance unbiased estimator of $\gamma(\theta)$ at $\theta = \theta_0$ if

$$E[g_*(X) g_*(X)^t | \theta_0] = \min_{g \in U_\gamma} E[g(X) g(X)^t | \theta_0].$$

Also it is easy to see that

$$U_\gamma = g + U_0, \quad \forall g \in U_\gamma.$$

Thus as an easy consequence of Theorem 2.2.1, we have the following generalization of the Lehmann-Scheffe theorem.

Corollary 2.2.1. With the same notation as above, $g_* \in U_\gamma$ is a locally minimum variance unbiased estimator of $\gamma(\theta)$ at $\theta = \theta_0$ iff

$$\langle g_*, h \rangle_{\theta_0} = E[g_*(X) h(X)^t | \theta_0] = 0,$$

$\forall h \in U_\gamma$.

Next we show that for any finite dimensional subspace in a generalized inner product space, the orthogonal projection always exists. In order to do this, the famous Gram-Schmidt orthogonalization procedure is used in generalized inner product spaces. We need another definition.

Definition 2.2.3. Let $(L, \langle \cdot, \cdot \rangle)$ be a generalized inner product space. A set of functions $\{h_i\}_{i=1}^n$ is said to be *linearly independent*, if for any set of $k \times k$ matrices $\{A_i\}_{i=1}^{n-1}$, defining

$$e_1 = h_1, \quad e_i = h_i - \sum_{j=1}^{i-1} A_j h_j, \quad i = 2, \dots, n,$$

$\{\langle e_i, e_i \rangle : i \in \{1, \dots, n\}\}$ are all invertible.

The following is the Gram-Schmidt orthogonalization procedure in generalized inner product spaces.

Proposition 2.2.1. If $\{h_i\}_{i=1}^n$ is linearly independent, let

$$e_1 = h_1, \quad e_2 = h_2 - \langle h_2, e_1 \rangle \langle e_1, e_1 \rangle^{-1} e_1,$$

$$e_k = h_k - \sum_{i=1}^{k-1} \langle h_k, e_i \rangle \langle e_i, e_i \rangle^{-1} e_i, \quad k \in \{2, \dots, n\}.$$

Then $\{e_i\}_{i=1}^n$ are orthogonal.

Proof. First note that

$$\langle e_2, e_1 \rangle = \langle h_2, e_1 \rangle - \langle h_2, e_1 \rangle = 0.$$

Now suppose that

$$\langle e_m, e_j \rangle = 0, \quad \forall \quad 1 \leq j \leq m \in \{2, \dots, k-1\}.$$

Then for all $j \in \{1, \dots, m\}$

$$\langle e_{m+1}, e_j \rangle = \langle h_{m+1}, e_j \rangle - \langle h_{m+1}, e_j \rangle = 0,$$

so that $\{e_i\}_{i=1}^n$ are orthogonal.

The above result is used to prove the existence of the orthogonal projection of every element of a generalized inner product space into a finite dimensional subspace.

Theorem 2.2.2. Let $(L, \langle \cdot, \cdot \rangle)$ be a generalized inner product space, and let L_0 be a finite dimensional subspace of L with linearly independent basis. Then for any $s \in L$, the orthogonal projection of s into L_0 always exists.

Proof. From Proposition 2.2.1, without loss of generality, we can assume that $\{h_1, \dots, h_m\}$ is an orthogonal basis for L_0 . Let

$$A_i = \langle s, h_i \rangle \langle h_i, h_i \rangle^{-1}, \quad i \in \{1, \dots, m\}.$$

We claim that the orthogonal projection of s into L_0 is

$$h_* = \sum_{i=1}^m A_i h_i.$$

To see this, for any $h = \sum_{i=1}^m b_j h_j \in L_0$,

$$\begin{aligned}
 & \langle s - h_*, h \rangle = \langle s, h \rangle - \langle h_*, h \rangle \\
 & = \sum_{j=1}^m \langle s, h_j \rangle b_j^t - \sum_{j=1}^m [\sum_{i=1}^m A_i \langle h_i, h_j \rangle] b_j^t \\
 & = \sum_{j=1}^m \langle s, h_j \rangle b_j^t - \sum_{j=1}^m A_j \langle h_j, h_j \rangle b_j^t \\
 & = \sum_{j=1}^m [\langle s, h_j \rangle - A_j \langle h_j, h_j \rangle] b_j^t \\
 & = 0.
 \end{aligned}$$

Now apply Theorem 2.2.1.

Theorem 2.2.2 will be used repeatedly in the subsequent sections for the derivation of optimal estimating functions.

Next we establish an abstract information inequality, which is fundamental to our later study. The motivation for the following definition will be clear from the subsequent section where we define information related to an estimating function.

Definition 2.2.4. Let $(L, \langle \cdot, \cdot \rangle)$ be a generalized inner product space, let $s \in L$ be a fixed element. For any $g \in L$, the information of g with respect to s is defined as

$$I_g = \langle g, s \rangle^t \langle g, g \rangle^- \langle g, s \rangle, \quad (2.2.5)$$

where “-” denotes a generalized inverse.

We shall also need the following theorem later.

Theorem 2.2.3. Let $(L, \langle \cdot, \cdot \rangle)$ be a generalized inner product space, and let L_0 be a linear subspace of L . For any $s \in L$, suppose g^* is the orthogonal projection of s into L_0 . Consider the function

$$I_g = \langle g, s \rangle^t \langle g, g \rangle^- \langle g, s \rangle.$$

Then

$$I_{g^*} - I_g$$

is n. n. d., for all $g \in L_0$.

Proof. Let $s = g_* + h$. Then using $\langle g, h \rangle = 0$,

$$\begin{aligned} I_g &= \langle g, s \rangle^t \langle g, g \rangle^- \langle g, s \rangle \\ &= \langle g, g_* \rangle^t \langle g, g \rangle^- \langle g, g_* \rangle, \end{aligned}$$

Also, using $\langle g_*, h \rangle = 0$,

$$\begin{aligned} I_{g_*} &= \langle g_*, s \rangle^t \langle g_*, g_* \rangle^- \langle g_*, s \rangle \\ &= \langle g_*, g_* \rangle. \end{aligned}$$

Now consider the matrix

$$\begin{bmatrix} \langle g_*, g_* \rangle & \langle g, g_* \rangle^t \\ \langle g, g_* \rangle & \langle g, g \rangle \end{bmatrix}.$$

For any k -dimensional vectors a and b , we have that

$$\begin{aligned} \begin{bmatrix} a^t & b^t \end{bmatrix} \begin{bmatrix} \langle g_*, g_* \rangle & \langle g, g_* \rangle^t \\ \langle g, g_* \rangle & \langle g, g \rangle \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} \\ = a^t \langle g_*, g_* \rangle a + 2a^t \langle g, g_* \rangle^t b + b^t \langle g, g \rangle b \\ = \langle a^t g_* + b^t g, a^t g_* + b^t g \rangle \geq 0. \end{aligned}$$

Thus

$$\begin{bmatrix} \langle g_*, g_* \rangle & \langle g, g_* \rangle^t \\ \langle g, g_* \rangle & \langle g, g \rangle \end{bmatrix}.$$

is n. n. d., which implies that

$$I_{g_*} - I_g = \langle g_*, g_* \rangle - \langle g, g_* \rangle^t \langle g, g \rangle^- \langle g, g_* \rangle$$

is n. n. d. The proof of the theorem is complete.

The following result will be used to establish the essential uniqueness of optimal estimating function.

Theorem 2.2.4. With the same notation as above, if $g^*, g \in L_0$, and g^* is the orthogonal projection of s into L_0 , and $\langle g^*, g^* \rangle, \langle g, g \rangle$ are invertible, then $I_{g^*} = I_g$ if and only if there exists an invertible matrix M such that

$$g^* = M g.$$

Proof. If. If $g^* = M g$, by straightforward calculation, we get

$$\begin{aligned} I_{g^*} &= \langle g^*, s \rangle^t \langle g^*, g^* \rangle^{-1} \langle g^*, s \rangle \\ &= \langle g, s \rangle^t M^t [M \langle g, g \rangle M^t]^{-1} M \langle g, s \rangle \\ &= \langle g, s \rangle^t \langle g, g \rangle^{-1} \langle g, s \rangle. \end{aligned}$$

Only if. If $I_{g^*} = I_g$, note that

$$I_{g^*} = \langle g^*, s \rangle^t \langle g^*, g^* \rangle^{-1} \langle g^*, s \rangle = \langle g^*, g^* \rangle,$$

and

$$I_g = \langle g, s \rangle^t \langle g, g \rangle^{-1} \langle g, s \rangle = \langle g, g^* \rangle^t \langle g, g \rangle^{-1} \langle g, g^* \rangle,$$

since g^* is the orthogonal projection of s . Then

$$0 = I_{g^*} - I_g = \langle g^*, g^* \rangle - \langle g, g^* \rangle^t \langle g, g \rangle^{-1} \langle g, g^* \rangle.$$

Let $M = \langle g, g^* \rangle^t \langle g, g \rangle^{-1}$, then it is easy to verify that

$$\begin{aligned} \langle g^* - M g, g^* - M g \rangle &= \langle g^*, g^* \rangle - M \langle g, g^* \rangle - \langle g^*, g \rangle M^t + M \langle g, g \rangle M^t \\ &= \langle g^*, g^* \rangle - \langle g, g^* \rangle^t \langle g, g \rangle^{-1} \langle g, g^* \rangle = 0, \end{aligned}$$

so that $g^* = M g$.

As an easy consequence of Theorems 2.2.2-2.2.4, we have the following corollary.

Corollary 2.2.2. Let $(L, \langle \cdot, \cdot \rangle)$ be a generalized inner product space, and let L_0 be a finite dimensional subspace of L with linearly independent basis. For all $s \in L$, and $g \in L_0$, let

$$I_g = \langle g, s \rangle^t \langle g, g \rangle^{-1} \langle g, s \rangle.$$

Then there exists $g^* \in L_0$ such that

$$I_{g^*} - I_g \quad (2.2.6)$$

is n. n. d., for all $g \in L_0$. Furthermore if $\langle g^*, g^* \rangle$ and $\langle g, g \rangle$ are invertible, then $I_{g^*} = I_g$ if and only if there exists an invertible matrix M such that

$$g_* = M g.$$

Proof. Since L_0 is a finite dimensional subspace of L with linearly independent basis, then by Theorem 2.2.2, for any $s \in L$, the orthogonal projection g^* of s into L_0 exists. The first part of the corollary now follows from Theorem 2.2.3. The second part of the corollary follows from Theorem 2.2.4.

2.3 Optimal Estimating Functions: A Geometric Approach

In this section, we will apply the results obtained in the previous section to the theory of estimating functions. We begin with the definition of unbiased estimating functions.

Let \mathcal{X} be a sample space and Θ be a k dimensional parameter space. A function

$$g : \mathcal{X} \times \Theta \longrightarrow R^k$$

is an unbiased estimating function if

$$E[g(X, \theta) | \theta] = 0, \quad \forall \theta \in \Theta.$$

An unbiased estimating function g is called regular if the following conditions hold:

(i) $d_{ij}(\theta) = E[\frac{\partial g_i}{\partial \theta_j} | \theta]$, ($1 \leq i, j \leq k$) exists;

(ii) $E[g(X, \theta) g(X, \theta)^t | \theta]$ is positive definite.

Let L denote the space of all regular unbiased estimating functions. For $g_1, g_2 \in L$, we define the family of generalized inner products of g_1, g_2 as

$$\langle g_1, g_2 \rangle_\theta = E[g_1(X, \theta) g_2(X, \theta)^t | \theta] \quad \forall \theta \in \Theta. \quad (2.3.7)$$

This family of generalized inner products will be used throughout this section without specific reference to it. Also we shall denote by s the score function of a parametric family of distributions. We assume also that the score vector is regular in the sense described in (i) and (ii).

Definition 2.3.1. With the same notation as above, let $(L, \langle \cdot, \cdot \rangle_\theta)$ be the family of generalized inner product spaces, and let L_0 be a subspace of L . For any $g \in L_0$, the information function of g is defined as follows

$$I_g(\theta) = E\left[\frac{\partial g}{\partial \theta}|\theta\right]^t \langle g, g \rangle_\theta^{-1} E\left[\frac{\partial g}{\partial \theta}|\theta\right] \quad (2.3.8)$$

An element $g^* \in L_0$ is said to be an optimal estimating function in L_0 if

$$I_{g^*}(\theta) - I_g(\theta)$$

is n. n. d., for all $g \in L_0$ and $\theta \in \Theta$.

Next we prove a key result which shows that definition (2.3.8) is indeed equivalent to definition (2.2.5) of the previous section.

In the rest of this section, unless otherwise stated, we shall assume the following regularity condition for unbiased estimating functions.

(\mathcal{R}). For any $g \in L$,

$$E\left[\frac{\partial g}{\partial \theta}|\theta\right] = -E[gs^t|\theta]. \quad (2.3.9)$$

Lemma 2.3.1. Under the regularity condition (\mathcal{R}), for any $g \in L$, the information matrix of g can be written as

$$I_g(\theta) = \langle g, s \rangle_\theta^t \langle g, g \rangle_\theta^{-1} \langle g, s \rangle_\theta,$$

where s is the score function.

Proof. The result follows easily since for any $g \in L$, use (2.3.9) to get

$$- \langle g, s \rangle_\theta = E\left[\frac{\partial g}{\partial \theta}|\theta\right].$$

Theorem 2.3.1. Let L_0 be a subspace of L . Assume that the orthogonal projection g^* of s into L_0 exists. Then

$$I_g(\theta) \leq I_{g^*}(\theta), \quad \forall \theta \in \Theta, \quad g \in L_0, \quad (2.3.10)$$

that is $g^* \in L_0$ is an optimal estimating function in L_0 . The optimal element in L_0 is unique in the following sense: if $g \in L_0$, then $I_g(\theta) = I_{g^*}(\theta)$, $\forall \theta \in \Theta$, if and only if there exists invertible matrix valued function $M : \Theta \rightarrow M_{k \times k}$ such that for any $\theta \in \Theta$,

$$g^*(X, \theta) = M(\theta) g(X, \theta), \quad (2.3.11)$$

with probability 1 with respect to P_θ .

Proof. The first part follows easily from Lemma 2.3.1 and Theorem 2.2.3. The second part follows from Theorem 2.2.4.

Note that if L_0 is a finite dimensional subspace of L , from Theorem 2.2.2, an orthogonal projection g^* of $s(\in L)$ into L_0 always exists, so that the conclusions given in (2.3.10) and (2.3.11) always hold. Also, in this case, Proposition 2.2.1 and Theorem 2.2.2 show how to construct optimal estimating functions.

In the remainder of this section, we shall see several applications of Theorem 2.3.1 for deriving optimal estimating functions in different contexts. We begin by generalizing a result of Godambe (1985) when the parameter space is multidimensional. Also we bring out more explicitly the characterization of optimal estimating functions in a more general framework than what is given in Theorem 1 of Godambe (1985). Let $\{X_1, X_2, \dots, X_n\}$ be a discrete stochastic process, $\Theta \subset R^k$ be an open set. Let h_i be a R^k valued function of X_1, \dots, X_i and θ to R^k , such that

$$E_{i-1}[h_i(X_1, \dots, X_i; \theta) | \theta] = 0, \quad (i = 1, \dots, n, \quad \theta \in \Theta), \quad (2.3.12)$$

where E_{i-1} denotes the conditional expectation conditioning on the first $i-1$ variables, namely, X_1, \dots, X_{i-1} . Let

$$L_0 = \{g : g = \sum_{i=1}^n A_{i-1} h_i\},$$

where A_{i-1} is a $M_{k \times k}$ valued function of X_1, \dots, X_{i-1} and θ , for all $i \in \{1, \dots, n\}$.

The following theorem generalizes the result of Godambe (1985).

Theorem 2.3.2. With the same notations as above, suppose h_i satisfies the regularity condition (\mathcal{R}) . Let

$$A_i^* = E_{i-1} \left[\frac{\partial h_i}{\partial \theta} | \theta \right]^t E_{i-1} [h_i h_i^t | \theta]^{-1} \quad \forall i \in \{1, 2, \dots, n\},$$

and

$$g^* = \sum_{i=1}^n A_i^* h_i.$$

Then the following conclusions hold:

- (a). g^* is the orthogonal projection of s into L_0 .
- (b). g^* is an optimal estimating function in L_0 , i. e.,

$$I_g(\theta) \leq I_{g^*}(\theta),$$

for all $g \in L_0$ and $\theta \in \Theta$.

(c). If $g \in L_0$ and $E[g g^t | \theta]$ is invertible, then $I_g(\theta) = I_{g^*}(\theta)$, for all $\theta \in \Theta$ if and only if there exists an invertible matrix function $M : \Theta \longrightarrow M_{k \times k}$ such that for any $\theta \in \Theta$,

$$g_*(X_1, \dots, X_n; \theta) = M(\theta) g(X_1, \dots, X_n; \theta),$$

with probability 1 with respect to P_θ .

Proof. (a). For any $g = \sum_{i=1}^n A_i h_i \in L_0$, $\theta \in \Theta$,

$$\begin{aligned} & \langle s - g^*, g \rangle_\theta = \langle s, g \rangle_\theta - \langle g^*, g \rangle_\theta \\ &= \sum_{i=1}^n E[s h_i^t A_i^t | \theta] - \sum_{i=1}^n \sum_{j=1}^n E[A_i^* h_i h_j^t A_j^t | \theta] \\ &= \sum_{i=1}^n E\{E_{i-1}[s h_i^t A_i^t | \theta]\} - \sum_{i=1}^n E[A_i^* h_i h_i^t A_i^t | \theta] \\ &\quad - \sum_{i < j} E[A_i^* h_i h_j^t A_j^t | \theta] - \sum_{i > j} E[A_i^* h_i h_j^t A_j^t | \theta] \end{aligned} \tag{2.3.13}$$

But for $i < j$,

$$\begin{aligned} E[A_i^* h_i h_j^t A_j^t | \theta] &= E\{E_{j-1}[A_i^* h_i h_j^t A_j^t | \theta] | \theta\} \\ &= E\{A_i^* h_i E_{j-1}[h_j^t A_j^t | \theta] | \theta\} = 0. \end{aligned}$$

Similarly, for $i > j$,

$$E[A_i^* h_i h_j^t A_j^t | \theta] = 0.$$

Thus from equation (2.3.13), we get

$$\begin{aligned} \langle s - g^*, g \rangle_\theta &= \sum_{i=1}^n E\{E_{i-1}[\frac{\partial h_i}{\partial \theta} | \theta]^t A_i^t | \theta\} - \\ &\quad \sum_{i=1}^n E\{A_i^* E_{i-1}[h_i h_i^t | \theta] A_i^t | \theta\} = 0. \end{aligned}$$

Hence g^* is the orthogonal projection of s into L_0 .

Parts (b) and (c) of the theorem follows easily from part (a) and Theorem 2.3.1.

A second application of Theorem 2.3.1 is to give a geometric formulation of a result of Godambe and Thompson (1989), who proved the existence of optimal estimating functions using mutually orthogonal estimating functions. What we show is that the optimal estimating function of Godambe and Thompson is indeed the orthogonal projection of the score function into an appropriate linear subspace.

To this end, let \mathcal{X} denote the sample space, $\theta = (\theta_1, \dots, \theta_m)$ be a vector of parameters, $h_j, j = 1, \dots, k$ be real functions on $\mathcal{X} \times \Theta$ such that

$$E[h_j(X, \theta) | \theta, \mathcal{X}_j] = 0, \quad \forall \theta \in \Theta, \quad j = 1, \dots, k,$$

where \mathcal{X}_j is a specified partition of $\mathcal{X}, j = 1, \dots, k$. We will denote

$$E[. | \theta, \mathcal{X}_j] = E_{(j)}[. | \theta].$$

Consider the class of estimating functions

$$L_0 = \{g : g = (g_1, \dots, g_m)\}$$

where

$$g_r = \sum_{j=1}^k q_{jr} h_j, \quad r = 1, \dots, m,$$

$q_{jr} : \mathcal{X} \times \Theta \longrightarrow R$ being measurable with respect to the partition \mathcal{X}_j for $j = 1, \dots, k, r = 1, \dots, m$.

Let

$$q_{jr}^* = \frac{E_{(j)}[\frac{\partial h_j}{\partial \theta_r} | \theta]}{E_{(j)}[h_j^2 | \theta]}, \quad (2.3.14)$$

for all $j = 1, \dots, k, r = 1, \dots, m$, and

$$g_r^* = \sum_{j=1}^k q_{jr}^* h_j, \quad r = 1, \dots, m.$$

The estimating functions $h_j, j = 1, \dots, k$ are said to be mutually orthogonal if

$$E_{(j)}[q_{jr}^* h_j q_{j'r'}^* h_{j'} | \theta] = 0, \quad \forall j \neq j', r, r' = 1, \dots, m. \quad (2.3.15)$$

Theorem 2.3.3. With the same notations as above, if $\{h_j\}_{j=1}^k$ are mutually orthogonal, then the following hold:

(a) g^* is the orthogonal projection of the score function s into L_0 .

(b) g^* is an optimal estimating function in L_0 .

(c). If $g \in L_0$, and $E[g g^t | \theta]$ is invertible, then $I_g(\theta) = I_{g^*}(\theta)$, $\forall \theta \in \Theta$ if and only if there exists an invertible matrix function $M : \Theta \longrightarrow M_{k \times k}$ such that for any $\theta \in \Theta$,

$$g^*(X; \theta) = M(\theta) g(X; \theta),$$

with probability 1 with respect to P_θ .

Proof. (1). We only need to show that, $\forall r \in \{1, \dots, m\}, g_r = \sum_{j=1}^k q_{jr} h_j$,

$$\langle s - g_r^*, g_r \rangle_\theta = 0, \quad \forall \theta \in \Theta$$

i.e.,

$$\langle s, g_r \rangle_\theta = \langle g_r^*, g_r \rangle_\theta, \quad \forall \theta \in \Theta.$$

But

$$\begin{aligned}
\langle g_r^*, g_r \rangle_\theta &= \sum_{j=1}^k \sum_{j'=1}^k E[q_{jr}^* h_j q_{j'r} h_{j'} | \theta] \\
&= \sum_{j=1}^k \sum_{j'=1}^k E\{q_{j'r}^{*-1} h_j q_{j'r} E_{(j)}[q_{jr}^* h_j q_{j'r} h_{j'} | \theta] | \theta\} \\
&= \sum_{j=1}^k E\{q_{jr}^* q_{jr} E_{(j)}[h_j^2 | \theta] | \theta\} \\
&= \sum_{j=1}^k E\{q_{jr} E_{(j)}[\frac{\partial h_j}{\partial \theta_r} | \theta] | \theta\}.
\end{aligned}$$

Also

$$\begin{aligned}
\langle s, g_r \rangle_\theta &= \sum_{j=1}^k E[q_{jr} s h_j | \theta] \\
&= \sum_{j=1}^k E\{q_{jr} E_{(j)}[s h_j | \theta] | \theta\} \\
&= \sum_{j=1}^k E\{q_{jr} E_{(j)}[\frac{\partial h_j}{\partial \theta_r} | \theta] | \theta\}.
\end{aligned}$$

Thus g^* is the orthogonal projection of the score function into L_0 .

Once again (b) and (c) follows from part (a) and Theorem 2.3.1. This completes the proof.

Note that part (b) of Theorem 2.3.3 is due to Godambe and Thompson (1989), while the other two parts are new. We repeat that this theorem provides a geometric formulation of optimal estimating functions in a finite dimensional subspace of estimating functions. Also through this approach, the characterization of optimal estimating function is very easy to establish.

Finally, we apply the above result to obtain optimal generalized estimating equations for multivariate data. Let \mathcal{X}_j denote the sample space for the j th subject, $\Theta \subset R^d$ be a subset with nonempty interior,

$$u_i : \mathcal{X}_j \times \Theta \longrightarrow R^{n_i}, \quad i = 1, \dots, k,$$

such that $E[u_i(X_i, \theta) | \theta] = 0, \forall \theta \in \Theta$. Suppose that conditional on θ , $\{u_i(X_i, \theta)\}_{i=1}^k$ are independent. Consider the estimating space

$$L_0 = \{\sum_{i=1}^k W_i(\theta) u_i(X_i, \theta)\}$$

where $W_i(\theta)$ is a $d \times n_i$ matrix, $i = 1, \dots, k$. Let

$$W_i^*(\theta) = E\left[\frac{\partial u_i}{\partial \theta}|\theta\right]^t [Var(u_i|\theta)]^{-1}, \quad i = 1, \dots, k,$$

$$g^* = \sum_{i=1}^k W_i^*(\theta) u_i(X_i, \theta).$$

Then we have the following result.

Theorem 2.3.4. With the same notations as above,

(a) g^* is the orthogonal projection of the score function into L_0 .

(b) g^* is an optimal estimating function in L_0 .

(c) If $g \in L_0$, and $E[g g^t|\theta]$ is invertible, then $I_g(\theta) = I_{g^*}(\theta)$, $\forall \theta \in \Theta$ if and only if there exists an invertible matrix function $M : \Theta \rightarrow M_{k \times k}$ such that for any $\theta \in \Theta$,

$$g^*(X; \theta) = M(\theta) g(X; \theta),$$

with probability 1 with respect to P_θ .

Proof. (a). We only need to show that $\forall g = \sum_{i=1}^k W_i(\theta) u_i(X_i, \theta)$,

$$\langle s, g \rangle_\theta = \langle g^*, g \rangle_\theta.$$

But

$$\begin{aligned} \langle g^*, g \rangle_\theta &= \sum_{i=1}^k \sum_{i'=1}^k W_i^* \langle u_i(X_i, \theta), u_{i'}(X_{i'}, \theta) \rangle_\theta W_{i'}^t \\ &= \sum_{i=1}^k W_i^* \langle u_i(X_i, \theta), u_i(X_i, \theta) \rangle_\theta W_i^t \\ &= \sum_{i=1}^k E\left[\frac{\partial u_i}{\partial \theta}|\theta\right]^t W_i^t; \end{aligned}$$

also

$$\begin{aligned} \langle s, g \rangle_\theta &= \sum_{i=1}^k \langle s, u_i \rangle_\theta W_i^t \\ &= \sum_{i=1}^k E\left[\frac{\partial u_i}{\partial \theta}|\theta\right]^t W_i^t. \end{aligned}$$

Thus g^* is the orthogonal projection of the score function into L_0 .

Parts (b) and (c) follows from part (a) and Theorem 2.3.1.

Note that by choosing the appropriate the functions of u_i , we can very easily get the generalized estimating equations introduced by Liang and Zeger (1986). For further information about generalized estimating equations, we refer to Liang, Zeger and Qaqish (1992).

2.4 Optimal Bayesian Estimating Functions

In this section, we study the geometry of estimating functions within a Bayesian framework. There are two basic approaches here. One formulation is based on the joint distribution of the data and prior, as introduced by Ferreira (1981, 1982). The second formulation, due to Ghosh (1990), is based on the posterior density. We shall study both and see how the notion of orthogonal projection can be brought within Bayesian formulation as well.

We begin with Ferreira's (1981, 1982) formulation. Let \mathcal{X} be the sample space, $\Theta \subset R^k$ be an open set, $p(x|\theta)$ be the conditional density of X given θ , and $\pi(\theta)$ be a prior density. Let $g : \mathcal{X} \times \Theta \longrightarrow R^k$ be a function such that

$$(1) \frac{\partial g}{\partial \theta} \text{ exists, } \forall \theta \in \Theta;$$

(2) $E[g(X, \theta)g(X, \theta)^t]$ is invertible, where E denotes expectation over the joint distribution of X and θ .

Let L denote the set of all functions $g : \mathcal{X} \times \Theta \longrightarrow R^k$ which satisfy (1) and (2) above. The generalized inner product on L is defined by

$$\langle f, g \rangle = E[f(X, \theta)g(X, \theta)^t], \quad \forall f, g \in L, \quad (2.4.16)$$

It is straightforward to verify that (2.4.16) is a generalized inner product on L .

The following calculation will be used to serve as a key connection between the formulation of Ferreira about optimal Bayesian estimating functions and our geometric

formulation. It also provides a geometric insight to the result of Ferreira. Throughout this section, we shall always assume that $p(X|\theta)$ and $\pi(\theta)$ are differentiable with respect to θ .

Lemma 2.4.1. Let $\pi(\theta|X)$ be the posterior density, and

$$s_j = \frac{\partial \log \pi(\theta|X)}{\partial \theta_j}, \quad \forall j \in \{1, \dots, k\},$$

and $g_i : \mathcal{X} \times \Theta \longrightarrow R$ be a function. Then

$$E[g_i s_j] = -E\left[\frac{\partial g_i}{\partial \theta_j}\right] + E\left\{\frac{\partial E[g_i|\theta]}{\partial \theta_j} + E[g_i|\theta] \frac{\partial \log \pi(\theta)}{\partial \theta_j}\right\}. \quad (2.4.17)$$

Proof.

$$\begin{aligned} E[g_i s_j] &= E\left\{E[g_i] \frac{\partial \log \pi(\theta|X)}{\partial \theta_j} | \theta\right\} \\ &= E\left\{E[g_i] \left(\frac{\partial \log p(X|\theta)}{\partial \theta_j} + \frac{\partial \log \pi(\theta)}{\partial \theta_j}\right) | \theta\right\} \\ &= E\left\{E[g_i] \left(\frac{\partial \log p(X|\theta)}{\partial \theta_j}\right) | \theta\right\} + E\left\{E[g_i] \frac{\partial \log \pi(\theta)}{\partial \theta_j} | \theta\right\} \\ &= E\left\{\frac{\partial E[g_i|\theta]}{\partial \theta_j}\right\} - E\left\{E\left[\frac{\partial g_i}{\partial \theta_j}\right]\right\} + E\left\{E[g_i|\theta] \frac{\partial \log \pi(\theta)}{\partial \theta_j}\right\} \\ &= -E\left[\frac{\partial g_i}{\partial \theta_j}\right] + E\left\{\frac{\partial E[g_i|\theta]}{\partial \theta_j} + E[g_i|\theta] \frac{\partial \log \pi(\theta)}{\partial \theta_j}\right\}. \end{aligned}$$

This completes the proof.

Note that if $E[g_i|\theta] = 0$, then

$$E[g_i s_j] = -E\left[\frac{\partial g_i}{\partial \theta_j}\right]; \quad (2.4.18)$$

also if g_i is only a function of θ , then

$$E[g_i s_j] = E\{E[g_i s_j | \theta]\} = E\left[g_i \frac{\partial \log \pi(\theta)}{\partial \theta_j}\right]. \quad (2.4.19)$$

Suppose now

$$B_{ij}(g) = E\left\{\frac{\partial E[g_i|\theta]}{\partial \theta_j} + E[g_i|\theta] \frac{\partial \log \pi(\theta)}{\partial \theta_j}\right\}, \quad (2.4.20)$$

for $i = 1, \dots, k, j = 1, \dots, k$, where $g = (g_1, \dots, g_k)$. Let $s = (s_1, \dots, s_k)$, using Lemma 2.4.1, then

$$\langle g, s \rangle = -((E[\frac{\partial g_i}{\partial \theta_j}] - B_{ij}(g))).$$

If $E[g|\theta] = 0$, from (2.4.17) and (2.4.18),

$$\langle g, s \rangle = -E[\frac{\partial g}{\partial \theta}]; \quad (2.4.21)$$

also if g is only a function of θ , then

$$\langle g, s \rangle = E[g (\frac{\partial \log \pi(\theta)}{\partial \theta})^t]. \quad (2.4.22)$$

Now by combining the previous theorem and the above lemma, we have the following result, which is a generalization of the main result due to Ferreira (1981, 1982) to the multidimensional case.

Theorem 2.4.1. For $g \in L$, let

$$M_g = E[g(X, \theta) g(X, \theta)^t], \quad (2.4.23)$$

then

$$((E[\frac{\partial g_i}{\partial \theta_j}] - B_{ij}(g)))^t M_g^{-1} ((E[\frac{\partial g_i}{\partial \theta_j}] - B_{ij}(g))) \leq M_s,$$

for all $g \in L$.

Proof. From the previous Lemma,

$$\langle g, s \rangle = -((E[\frac{\partial g_i}{\partial \theta_j}] - B_{ij}(g))).$$

Also $M_s = \langle s, s \rangle^t \langle s, s \rangle^{-1} \langle s, s \rangle$. Thus the result follows easily from Theorem 2.2.3.

Note that if $k = 1$, the above theorem reduces to the result proved by Ferreira (1981, 1982).

For any $g \in L$, let

$$I_g = ((E[\frac{\partial g_i}{\partial \theta_j}] - B_{ij}(g)))^t M_g^{-1} ((E[\frac{\partial g_i}{\partial \theta_j}] - B_{ij}(g))). \quad (2.4.24)$$

In the definition of I_g , $((E[\frac{\partial g_i}{\partial \theta_j}] - B_{ij}(g)))$ is a measure of sensitivity of g , and M_g is a measure of variability of g . Thus, analogous to the frequentist case, the following definition seems to be appropriate about optimal estimating function in the Bayesian framework.

Definition. If L_0 is a subspace of L , and $g^* \in L_0$, g^* is called an optimal Bayesian estimating function in L_0 , if for any $g \in L$,

$$I_g \leq I_{g^*}.$$

Next we prove an optimality result about Bayesian estimating functions in this formulation.

Theorem 2.4.2. With the same notation as above, the generalized inner product on L is defined by (2.4.16), and let L_0 be a subspace of L , if g^* is the orthogonal projection of s into L_0 , then we have that

- (1) g^* is an optimal Bayesian estimating function in L_0 ;
- (2) the optimal Bayesian estimating function in L_0 is unique in the following sense: for any $g \in L_0$, $I_g = I_{g^*}$ if and only if there exists an invertible $k \times k$ matrix M such that $g^* = M g$.

Proof. (1). From Theorem 2.2.3,

$$\langle g, s \rangle^t \langle g, g \rangle^{-1} \langle g, s \rangle \leq \langle g^*, s \rangle^t \langle g^*, g^* \rangle^{-1} \langle g^*, s \rangle,$$

for all $g \in L_0$. But from Lemma 2.4.1,

$$I_g = \langle g, s \rangle^t \langle g, g \rangle^{-1} \langle g, s \rangle,$$

for any $g \in L_0$. Thus the result follows easily.

(2) follows easily from Theorem 2.2.4.

Next we apply Theorem 2.4.2 to a case where L_0 is a finite dimensional subspace of L , with linearly independent basis.

Let $\{u_i(X_i, \theta)\}_{i=1}^K$ be a family of $n_i \times 1$ vectors of parameteric functions and $v(\theta)$ be a $m \times 1$ vector such that

- (1). For fixed $\theta \in \Omega$, $u_i(\cdot, \theta) : \mathcal{X} \rightarrow R^{n_i}$ is measurable;
- (2). $v : \Theta \rightarrow R^m$ is measurable;
- (3). $E[u_i|\theta] = 0$, and $E[v] = 0$;
- (4). Conditional on θ , $\{u_i(X_i, \theta)\}_{i=1}^K$ are independent.

Consider the space of estimating functions of the form

$$L_0 = \{\Sigma_{i=1}^K [W_i(\theta) u_i(X_i, \theta_i)] + Q v(\theta)\},$$

where for any $\theta \in \Theta$, $W_i(\theta)$ is a $p \times n_i$ matrix, for all $i \in \{1, \dots, K\}$, and Q is a $p \times m$ matrix.

Theorem 2.4.3. With the same notation as above, let

$$W_i^*(\theta) = E\left[\frac{\partial u_i}{\partial \theta}|\theta\right]^t (E[Var(u_i|\theta)])^{-1}, \quad Q^* = E[v(\theta)] \left(\frac{\partial \log \pi}{\partial \theta}\right)^t (E[v(\theta) v(\theta)^t])^{-1},$$

and

$$g^* = \Sigma_{i=1}^K (W_i^*(\theta) u_i(X_i, \theta)) + Q^* v(\theta).$$

Then

- (a) g^* is the orthogonal projection of s into L_0 ;
- (b) g^* is an optimal Bayesian estimating function in L_0 ;

(c) optimal Bayesian estimating function in L_0 is unique in the following sense: if $g \in L_0$, then $I_g = I_{g^*}$ if and only if there exists an invertible matrix M such that

$$g^*(X_1, \dots, X_K; \theta) = M g(X_1, \dots, X_K; \theta),$$

with probability 1 with respect to the joint distribution of the X_i and θ .

Proof. (a). For any $g = \Sigma_{i=1}^K (W_i(\theta) u_i(X_i, \theta)) + Q v(\theta)$,

$$\langle s - g^*, g \rangle = \langle s, g \rangle - \langle g^*, g \rangle.$$

But

$$\begin{aligned} \langle s, g \rangle &= \Sigma_{i=1}^K E\{E[s u_i(X_i, \theta)^t | \theta] W_i(\theta)^t\} + E[s v(\theta)^t] Q^t \\ &= -\Sigma_{i=1}^K E\{E\left[\frac{\partial u_i(X_i, \theta)}{\partial \theta} | \theta\right]^t W_i(\theta)^t\} + E[v(\theta) \left(\frac{\partial \log \pi}{\partial \theta}\right)^t]^t Q^t, \end{aligned}$$

and

$$\begin{aligned} \langle g^*, g \rangle &= \Sigma_{i=1}^K E\{W_i^*(\theta) E[u_i(X_i, \theta) u_i(X_i, \theta)^t | \theta] W_i(\theta)^t\} + Q^* E[v(\theta) v(\theta)^t] Q^t \\ &= -\Sigma_{i=1}^K E\{E\left[\frac{\partial u_i(X_i, \theta)}{\partial \theta} | \theta\right]^t W_i(\theta)^t\} + E[v(\theta) \left(\frac{\partial \log \pi}{\partial \theta}\right)^t]^t Q^t. \end{aligned}$$

Thus by Theorem 2.2.1, g^* is the orthogonal projection of s into L_0 .

Parts (b) and (c) follows from (a) and Theorem 2.4.2.

Next we turn to the formulation of Bayesian estimating functions introduced by Ghosh (1990). In this formulation, the parameter space is assumed to have the form $\Theta = (a_1, b_1) \times \dots \times (a_k, b_k)$. We start with a result which is very similar to Lemma 2.4.1.

Lemma 2.4.2. Let $\pi(\theta|X)$ be the posterior density, and $s_j = \frac{\partial \log \pi(\theta|X)}{\partial \theta_j}$, $j = 1, \dots, k$, and $g_i : \mathcal{X} \times \Theta \longrightarrow R$ be a function with suitable regularity condition, then

$$E[g_i s_j | X] = -E\left[\frac{\partial g_i}{\partial \theta_j} | X\right] + E[B_j(g_i) | X],$$

where

$$B_j(g_i) = \lim_{\theta_j \rightarrow b_j^-} g_i(X, \theta) \pi(\theta|X) - \lim_{\theta_j \rightarrow a_j^+} g_i(X, \theta) \pi(\theta|X).$$

Proof. Note that

$$\begin{aligned} E[g_i s_j|X] &= \int_{\Theta} g_i \frac{\partial \pi(\theta|X)}{\partial \theta_j} d\theta \\ &= E[B_j(g_i)|X] - E\left[\frac{\partial g_i}{\partial \theta_j}|X\right]. \end{aligned}$$

Next the definition about posterior estimating functions is introduced. A function $g : \Theta \times \mathcal{X} \longrightarrow R^k$ is called a posterior unbiased estimating function (PUEF) if

$$E[g(\theta, X)|X] = 0, \quad (2.4.25)$$

$$B_j(g_i) = 0, \quad \forall x \in \mathcal{X}, \quad i, j \in \{1, \dots, k\}. \quad (2.4.26)$$

Actually, all we require is that

$$E[B_j(g_i)|X] = 0, \quad \forall x \in \mathcal{X}, \quad i, j \in \{1, \dots, k\}.$$

Let L be the space consists of all functions $g : \Theta \times \mathcal{X} \longrightarrow R^k$, which is PUEF and $E[g g^t|X]$ is invertible. A family of generalized inner products on L is defined as follows: for any $f, g \in L$, and $x \in \mathcal{X}$,

$$\langle f, g \rangle_x = E[f(\theta, X) g(\theta, X)^t | X = x]. \quad (2.4.27)$$

If the score function $s \in L$, then from Lemma 2.4.2,

$$\langle g, s \rangle_x = -\left(E\left[\frac{\partial g_i}{\partial \theta_j} | X = x\right]\right).$$

Next for every $g \in L, x \in \mathcal{X}$, define

$$I_g(x) = \left(E\left[\frac{\partial g_i}{\partial \theta_j} | X = x\right]\right)^t \left(E[g(\theta, X) g(\theta, X)^t | X = x]\right)^{-1} \left(E\left[\frac{\partial g_i}{\partial \theta_j} | X = x\right]\right). \quad (2.4.28)$$

Let L_0 be a subspace of L ; $g^* \in L_0$ is said to be an optimal element in L_0 if

$$I_g(x) \leq I_{g^*}(x),$$

for all $g \in L_0$, and $x \in \mathcal{X}$.

The following result now follow very easily.

Theorem 2.4.4. With the same notation as above, suppose that the orthogonal projection g^* of s into L_0 exists with respect to the generalized inner products. Then g^* is optimal in L_0 , that is, for all $g \in L_0$, we have that

$$I_g(x) \leq I_{g^*}(x),$$

$\forall x \in \mathcal{X}$. Furthermore, the optimal element in L_0 is unique in the following sense: if $g \in L_0$, then $I_g(x) = I_{g^*}(x)$, $\forall x \in \mathcal{X}$ if and only if there exists an invertible matrix valued function $M : \mathcal{X} \rightarrow M_{k \times k}$ such that

$$g(\theta; x) = M(x) g^*(\theta; x).$$

Proof. The first part of the theorem is a consequence of Theorem 2.2.3, and the second part is a consequence of Theorem 2.2.4.

Note that if $s \in L_0$, then s is an optimal estimating function.

As a corollary of Theorem 2.4.4, we have the following generalization of a result due to Godambe (1994) about optimal estimating functions to multi-dimensional parameter space.

Corollary 2.4.1. If $g^* \in L_0$ is the orthogonal projection of s into L_0 , then

- (a) $I_g(x) \leq I_{g^*}(x)$, for all $g \in L_0$ and $x \in \mathcal{X}$;
- (b) $E[(g^* - s)(g^* - s)^t | x] \leq E[(g - s)(g - s)^t | x]$, for all $g \in L_0$, and $x \in \mathcal{X}$.

Note that it is easy to see that if the parameter space is one dimensional, then (a) is equivalent to $\text{corr}\{g^*, s|x\}^2 \geq \text{corr}\{g, s|x\}^2$, for all $g \in L_0$ and $x \in \mathcal{X}$. This is the result proved by Godambe (1994).

2.5 Orthogonal Decomposition and Information Inequality

In this section, we give a geometric intuition of some information inequalities. Let us start with one of the main results of this section.

Theorem 2.5.1. Suppose L_0 is a subspace of L , and for all $g \in L$, let g_0 be the orthogonal projection of g into L_0 . Also, let s be the score function.

(i). If $\langle g - g_0, s \rangle_\theta = 0$, $\forall \theta \in \Theta$, then

$$I_g(\theta) \leq I_{g_0}(\theta), \quad \forall \theta \in \Theta.$$

(ii). If $\langle g_0, s \rangle_\theta = 0$, $\forall \theta \in \Theta$, then

$$I_g(\theta) \leq I_{g-g_0}(\theta), \quad \forall \theta \in \Theta.$$

Proof. Note that

$$I_g(\theta) = \langle g, s \rangle_\theta^t \langle g, g \rangle_\theta^{-1} \langle g, s \rangle_\theta, \quad \forall \theta \in \Theta.$$

(i). If $\langle g - g_0, s \rangle_\theta = 0$, $\forall \theta \in \Theta$, then $\langle g, s \rangle_\theta = \langle g_0, s \rangle_\theta$, $\forall \theta \in \Theta$. Also,

$$\langle g, g \rangle_\theta = \langle g_0, g_0 \rangle_\theta + \langle g - g_0, g - g_0 \rangle_\theta$$

$$\geq \langle g_0, g_0 \rangle_\theta, \quad \forall \theta \in \Theta.$$

Thus

$$I_g(\theta) \leq I_{g_0}(\theta), \quad \forall \theta \in \Theta.$$

(ii). If $\langle g_0, s \rangle_\theta = 0$, $\forall \theta \in \Theta$, then $\langle g - g_0, s \rangle_\theta = \langle g, s \rangle_\theta$, $\forall \theta \in \Theta$, and

$$\langle g, g \rangle_\theta = \langle g_0, g_0 \rangle_\theta + \langle g - g_0, g - g_0 \rangle_\theta$$

$$\geq \langle g - g_0, g - g_0 \rangle_\theta, \quad \forall \theta \in \Theta.$$

Thus

$$I_g(\theta) \leq I_{g-g_0}(\theta), \quad \forall \theta \in \Theta.$$

As an application of the previous result, we have the following:

Corollary 2.5.1. If T is a statistic, $\forall g \in L$, let $g_0 = E[g(X, \theta)|T]$. Then

(1) if T is sufficient, then

$$I_g(\theta) \leq I_{g_0}(\theta), \quad \forall \theta \in \Theta;$$

(2) if T is ancillary, then

$$I_g(\theta) \leq I_{g-g_0}(\theta), \quad \forall \theta \in \Theta.$$

Proof. (1). If T is sufficient, using the factorization theorem, we have that

$$\langle g - g_0, s \rangle_\theta = 0, \quad \forall \theta \in \Theta,$$

since the score function is only a function of T . The result follows now from part (i) of the previous theorem.

(2). If T is ancillary, then $\forall \theta \in \Theta$,

$$\begin{aligned} \langle g_0, s \rangle_\theta &= E[g_0 \left(\frac{\partial \log f_X}{\partial \theta} \right)^t | \theta] \\ &= E[g_0 f_T \left(\frac{\partial \log f_{X|T}}{\partial \theta} \right)^t | \theta] \\ &= E\{g_0 f_T E[(\frac{\partial \log f_{X|T}}{\partial \theta})^t | \theta, T] | \theta\} = 0, \end{aligned}$$

where f_X and $f_{X|T}$ are the marginal and conditional pdf's of X respectively, since the conditional score function has zero expectation with respect to the conditional density.

Thus both results follow from the previous theorem easily.

Next we study the information decomposition for information unbiased estimating functions. Let us start with the definition introduced by Lindsay (1982). Let g be an estimating function. Then g is called information unbiased if

$$\langle g, s \rangle_\theta = \langle g, g \rangle_\theta, \quad \forall \theta \in \Theta,$$

i. e. , g and $s - g$ are orthogonal, where s is the score function.

The main result on information unbiased estimating function is given in the following theorem.

Theorem 2.5.2. Let T be a statistic such that the marginal and conditional densities of T satisfy the usual regularity conditions. For all $g \in L$, let $g_0 = E[g|T, \theta]$, $h = g - E[g|T, \theta]$. Suppose that g is information unbiased. Then

- (1) if h is information unbiased with respect to $f_{X|T}$, then g_0 is information unbiased with respect to f_T ;
- (2) if g_0 is information unbiased with respect to f_T , then h is information unbiased with respect to $f_{X|T}$;
- (3) if at least one of (1) or (2) holds, then

$$I_g(\theta) = I_{g_0}(\theta) + I_h(\theta), \quad \forall \theta \in \Theta.$$

Proof. Let s_X denote the score function of X . Then we have that

$$s_X = s_T + s_{X|T},$$

and

$$\begin{aligned} < g, s_X - g >_{\theta} = E[g (s_X - g)^t | \theta] \\ &= E[(g_0 + h) (s_T - g_0 + s_{X|T} - h)^t | \theta] \\ &= < g_0, s_T - g_0 >_{\theta} + < h, s_{X|T} - h >_{\theta} + E[g_0 (s_{X|T} - h)^t | \theta] + E[h (s_T - g_0)^t | \theta]. \end{aligned}$$

But

$$E[g_0 (s_{X|T} - h)^t | \theta] = E\{g_0 E[(s_{X|T} - h)^t | T, \theta]\} = 0,$$

since $E[h|T, \theta] = 0, \forall \theta \in \Theta$, and

$$E[h (s_T - g_0)^t | \theta] = E\{E[h|T, \theta](s_T - g_0)^t | \theta\} = 0,$$

since $E[h|T, \theta] = 0, \forall \theta \in \Theta$. So

$$< g, s_X - g >_{\theta} = < g_0, s_T - g_0 >_{\theta} + < h, s_{X|T} - h >_{\theta}, \quad \forall \theta \in \Theta. \quad (2.5.29)$$

(1) and (2) follows from the above equality.

(3). First note that if g is information unbiased, then

$$I_g(\theta) = \langle g, g \rangle_\theta, \quad \forall \theta \in \Theta.$$

Also because $g = g_0 + h$, and g_0, h are orthogonal, so

$$\langle g, g \rangle_\theta = \langle g_0, g_0 \rangle_\theta + \langle h, h \rangle_\theta, \quad \forall \theta \in \Theta,$$

this implies that

$$I_g(\theta) = I_{g_0}(\theta) + I_h(\theta), \quad \forall \theta \in \Theta.$$

As an easy consequence of the above theorem, we have the following result proved by Bhapkar (1989, 1991a).

Corollary 2.5.2. With the same notation as the above theorem,

$$I_{s_X}(\theta) = I_{s_T}(\theta) + I_{s_{X|T}}(\theta), \quad \forall \theta \in \Theta.$$

Proof. The result follows by noting that under the usual regularity conditions, s_T and $s_{X|T}$ are information unbiased.

Note that if T is sufficient, then

$$I_{s_X}(\theta) = I_{s_T}(\theta), \quad \forall \theta \in \Theta.$$

If T is ancillary, then

$$I_{s_X}(\theta) = I_{s_{X|T}}(\theta), \quad \forall \theta \in \Theta.$$

2.6 Orthogonal Decomposition for Estimating Functions

In this section, we prove a Hoeffding type decomposition for estimating functions, revealing the geometric nature of the Hoeffding decomposition for U statistics.

Let X_1, \dots, X_n be an independent random variables, \mathcal{X}_i be the sample space for $X_i, i = 1, \dots, n, \Theta$ be the parameter space. A function

$$h : \mathcal{X}_1 \times \dots \times \mathcal{X}_n \times \Theta \longrightarrow R^k,$$

is called an unbiased estimating function if

$$E[h(X_1, \dots, X_n; \theta)|\theta] = 0, \quad \forall \theta \in \Theta.$$

Let \mathcal{E} consist of all unbiased functions $h : \mathcal{X}_1 \times \dots \times \mathcal{X}_n \times \Theta \longrightarrow R^k$, such that

$$E[h(X_1, \dots, X_n; \theta) \ h(X_1, \dots, X_n; \theta)^t | \theta]$$

is well defined for all $\theta \in \Theta$.

For $h_1, h_2 \in \mathcal{E}$, define a family of generalized inner product of h_1, h_2 as

$$\langle h_1, h_2 \rangle_\theta = E[h_1 \ h_2^t | \theta].$$

Then $(\mathcal{E}, \langle \cdot, \cdot \rangle_\theta)$ is a generalized inner product space for all $\theta \in \Theta$.

For $m \leq n$, let \mathcal{E}_m be a linear span of the functions of the form

$$h : \mathcal{X}_{i_1} \times \dots \times \mathcal{X}_{i_m} \times \Theta \longrightarrow R^k,$$

which satisfies

$$E[h(X_{i_1}, \dots, X_{i_m}; \theta)|\theta] = 0,$$

and $\langle h, h \rangle_\theta$ is well defined for all $\theta \in \Theta$, where $\{i_1, \dots, i_m\} \subset \{1, \dots, n\}$.

If $m_1 \leq m_2 \leq n$, then \mathcal{E}_{m_1} can be regarded as a subspace of \mathcal{E}_{m_2} in the obvious fashion. Now a natural question is: $\forall h \in \mathcal{E}$, $m \leq n$, does the orthogonal projection of h into \mathcal{E}_m exist? If it does, how do we find it? The answer to the above question is affirmative, and it turns out that the answer is very closely related to the Hoeffding type decomposition for U statistics. Before proving the main result of this section, let us introduce some further notation to simplify our presentation. For all $I = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}$, where $i_1 < \dots < i_m$, let

$$f(X_I) = f(X_{i_1}, \dots, X_{i_m}), \quad E[g|X_I] = E[g|X_{i_1}, \dots, X_{i_m}].$$

$$\mathcal{I}_k = \{(i_1, \dots, i_k) : 1 \leq i_1 < \dots < i_k \leq n\}, \quad \forall 1 \leq k \leq n.$$

Now we return to the orthogonal decomposition of estimating functions. Let $h \in \mathcal{E}$; $\forall I = \{i_1, \dots, i_m\} \subset \{1, \dots, n\}, 1 \leq k \leq n$, let

$$g_I(X_I) = E[h|X_I] - \sum_{J \subset I, J \neq I} E[h|X_J], \quad (2.6.30)$$

$\forall 1 \leq k \leq n$, let

$$h_k = \sum_{I \in \mathcal{I}_k} g_I(X_I). \quad (2.6.31)$$

Then we have the following result.

Theorem 2.6.1. Let \mathcal{E} and \mathcal{E}_m be the generalized inner product space defined as above, where $m \leq n$. Then $\forall h \in \mathcal{E}$, the orthogonal projection of h into \mathcal{E}_m exists, and is given by

$$\sum_{i=1}^m h_i,$$

where $h_i (1 \leq i \leq m)$ are defined as above.

Proof. We are only going to show that h_1, h_2 are the orthogonal projections of h into \mathcal{E}_1 and \mathcal{E}_2 respectively. The rest can be proved similarly.

(1). $\sum_{i=1}^n E[h|X_i]$ is the orthogonal projection of h into \mathcal{E}_1 . In fact, $\forall \sum_{j=1}^n g_j(X_j) \in \mathcal{E}_1$, we have

$$\begin{aligned} & \langle h - \sum_{i=1}^n E[h|X_i], \sum_{j=1}^n g_j(X_j) \rangle_{\theta} \\ &= \sum_{j=1}^n \langle h, g_j(X_j) \rangle_{\theta} - \sum_{i=1}^n \sum_{j=1}^n \langle E[h|X_i], g_j(X_j) \rangle_{\theta} \\ &= \sum_{j=1}^n \langle E[h|X_j], g_j(X_j) \rangle_{\theta} - \sum_{j=1}^n \langle E[h|X_j], g_j(X_j) \rangle_{\theta} \\ &= 0, \end{aligned}$$

since $\{X_i\}_{i=1}^n$ are independent. So by Theorem 2.2.1, we know that $\sum_{i=1}^n E[h|X_i]$ is the orthogonal projection of h into \mathcal{E}_1 .

(2). Next we show that

$$\sum_{i_1 < i_2} \{E[h|X_{i_1}, X_{i_2}] - E[h|X_{i_1}] - E[h|X_{i_2}]\} + \sum_{i=1}^n E[h|X_i]$$

is the orthogonal projection of h into \mathcal{E}_2 . If, $\forall \Sigma_{j_1 < j_2} g_{j_1, j_2}(X_{j_1}, X_{j_2}) \in \mathcal{E}_2$,

$$\begin{aligned}
& < h - h_2 - h_1, \Sigma_{j_1 < j_2} g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} \\
&= \Sigma_{j_1 < j_2} < h, g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} - \Sigma_{j_1 < j_2} < h_2, g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} - \\
&\quad \Sigma_{j_1 < j_2} < h_1, g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} \\
&= \Sigma_{j_1 < j_2} < E[h|X_{j_1}, X_{j_2}], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} - \\
&\Sigma_{j_1 < j_2} \Sigma_{i_1 < i_2} \{ < E[h|X_{i_1}, X_{i_2}] - E[h|X_{i_1}] - E[h|X_{i_2}], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} - \\
&\quad \Sigma_{j_1 < j_2} \Sigma_{i=1}^n < E[h|X_i], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} \} \tag{2.6.32}
\end{aligned}$$

Note that if $i_1 \neq j_1$ and $i_2 \neq j_2$, then

$$< E[h|X_{i_1}, X_{i_2}] - E[h|X_{i_1}] - E[h|X_{i_2}], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} = 0,$$

by the independence of $\{X_i\}_{i=1}^n$. If $i_1 = j_1$ and $i_2 \neq j_2$ or $i_1 \neq j_1$ and $i_2 = j_2$, then

$$< E[h|X_{i_1}, X_{i_2}] - E[h|X_{i_1}] - E[h|X_{i_2}], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} = 0.$$

Also if $i \neq j_1$ and $i \neq j_2$, then

$$< E[h|X_i], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} = 0.$$

Thus from the above equation, we get that

$$\begin{aligned}
& < h - h_2 - h_1, \Sigma_{j_1 < j_2} g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} \\
&= \Sigma_{j_1 < j_2} \{ < E[h|X_{j_1}] + E[h|X_{j_2}], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} \} - \\
&\quad \Sigma_{j_1 < j_2} \Sigma_{i \in j_1, j_2} < E[h|X_i], g_{j_1, j_2}(X_{j_1}, X_{j_2}) >_{\theta} \\
&= 0.
\end{aligned}$$

Hence $h_2 + h_1$ is the orthogonal projection of h into \mathcal{E}_2 .

As a consequence of the above theorem, we have the following result.

Corollary 2.6.1. Use the same notation as above, then $\{h_1, \dots, h_n\}$ are orthogonal to each other.

Proof. For all $1 \leq m_1 < m_2 \leq n$, since $\sum_{i=1}^{m_2-1} h_i$ and $\sum_{i=1}^{m_2} h_i$ are the orthogonal projections of h into \mathcal{E}_{m_2-1} and \mathcal{E}_{m_2} respectively, thus

$$h_{m_2} = \sum_{i=1}^{m_2} h_i - \sum_{i=1}^{m_2-1} h_i$$

is orthogonal to \mathcal{E}_{m_2-1} . Hence h_{m_2} is orthogonal to $\{h_1, \dots, h_{m_2-1}\}$, since

$$\{h_1, \dots, h_{m_2-1}\} \subset \mathcal{E}_{m_2-1}.$$

Theorem 15 generalizes the ANOVA decomposition for statistics proved by Efron and Stein (1981), to the estimating function case.

Also as another consequence of the orthogonal decomposition theorem, we have the following variance decomposition result.

Corollary 2.6.2. Use the same notation as above, we have that

$$Var_{\theta}(h) = \sum_{i=1}^n Var_{\theta}(h_i).$$

CHAPTER 3

THE GEOMETRY OF ESTIMATING FUNCTIONS II

3.1 Introduction

In Chapter 2, the notion of generalized inner product spaces is introduced to study optimal estimating functions without nuisance parameters. It was shown that the orthogonal projection of the *score function* into a linear subspace of estimating functions was optimal in that subspace, and a general method for the construction of such orthogonal projections was also given. As applications, both frequentist and Bayesian optimal estimating functions were found including as special cases some of the frequentist and Bayesian results derived earlier.

In this chapter, we extend the results of the previous chapter to find optimal estimating functions in the presence of nuisance parameters. First in Section 3.2, we derive some simple extension of the basic geometric results of Chapter 2. Next, in Section 3.3, we derive a general result on global optimal estimating functions which extends the results of Godambe and Thompson (1974) and Godambe (1976) to the multiparameter case. The general result is also used to study the geometry of conditional and marginal inference including as special cases some of the results of Bhapkar (1989, 1991).

In Section 3.4, a general result on locally optimal estimating functions is found, and is used to generalize (1) Lindsay's (1982) result on the local optimality of conditional score functions, (2) Godambe's (1985) result on estimation for stochastic processes, and (3) Murphy and Li's (1995) result on projected partial likelihood.

Finally, in Section 3.5, we derive optimal conditional estimating functions. As an application, we generalize the results of Godambe and Thompson (1989) in the presence of nuisance parameters.

3.2 Properties of Orthogonal Projections

The following result demonstrates that the operation of orthogonal projection is compatible with linear operations in a generalized inner product space.

Proposition 3.2.1. Let $(L, \langle \cdot, \cdot \rangle)$ be a generalized inner product space, and let L_0 be a subspace of L . If $s_1, s_2 \in L$, and the orthogonal projection g_i of s_i into L_0 exists for $i = 1, 2$, then

- (i) $g_1 + g_2$ is the orthogonal projection of $s_1 + s_2$ into L_0 ;
- (ii) for any matrix N , $N g_1$ is the orthogonal projection of $N s_1$ into $N L_0$.

Proof. From Theorem 2.2.1, g^* is the orthogonal projection of s into L_0 if and only if

$$\langle s - g^*, g \rangle = 0,$$

for all $g \in L_0$.

- (i) For any $g \in L_0$,

$$\langle s_1 + s_2 - g_1 - g_2, g \rangle = \langle s_1 - g_1, g \rangle + \langle s_2 - g_2, g \rangle = 0;$$

so $g_1 + g_2$ is the orthogonal projection of $s_1 + s_2$ into L_0 .

- (ii) For any $g \in N L_0$,

$$\langle N s_1 - N g_1, g \rangle = \langle N (s_1 - g_1), g \rangle$$

$$= N \langle s_1 - g_1, g \rangle = 0;$$

so $N g_1$ is the orthogonal projection of $N s_1$ into L_0 .

The following result is a slight generalization of Theorem 2.2.3

Theorem 3.2.1. Let $(L, \langle \cdot, \cdot \rangle)$ be a generalized inner product space, and let L_0 be a subspace of L . For any fixed $s \in L$, and any invertible matrix N , suppose that the orthogonal projection g^* of $N s$ into L_0 exists. Consider the function

$$I_g = \langle g, s \rangle^t \langle g, g \rangle^{-1} \langle g, s \rangle.$$

Then

$$I_{g^*} - I_g$$

is non negative definite, for all $g \in L_0$.

Proof. Since g^* is the orthogonal projection of $N s$ into L_0 and N is invertible, $N^{-1} g^*$ is the orthogonal projection of s into $N^{-1} L_0$, from part (ii) of Proposition 1. Hence, from part (b) of Theorem 2.2.3,

$$I_{N^{-1} g^*} - I_{N^{-1} g} \tag{3.2.1}$$

is non negative definite for all $g \in L_0$. But for any g ,

$$\begin{aligned} I_{N^{-1} g} &= \langle N^{-1} g, s \rangle^t \langle N^{-1} g, N^{-1} g \rangle^{-1} \langle N^{-1} g, s \rangle \\ &= \langle g, s \rangle^t (N^t)^{-1} [N^{-1} \langle g, g \rangle (N^t)^{-1}]^{-1} N^{-1} \langle g, s \rangle \\ &= \langle g, s \rangle^t \langle g, g \rangle^{-1} \langle g, s \rangle = I_g. \end{aligned}$$

Hence, the result follows from (3.2.1).

3.3 Global Optimality of Estimating Functions

In this section, a general result about global optimality of estimating functions in the presence of nuisance parameters will be proved. As easy consequences of this result, some results of Godambe and Thompson (1974), and Godambe (1976) are found. Further, the geometry of conditional and marginal inferences will be explored.

3.3.1 The General Result

Suppose \mathcal{X} is a sample space, $\Theta = \Theta_1 \times \Theta_2$ is the parameter space, with $\Theta_i \subset R^{d_i}$ ($i = 1, 2$). Let $\theta = (\theta_1, \theta_2)$, $\theta_1 \in \Theta_1$ be the parameter of interest, and $\theta_2 \in \Theta_2$ be the nuisance parameter. Consider the function

$$g : \mathcal{X} \times \Theta_1 \longrightarrow R^{d_1},$$

where g satisfies the following conditions:

(I) $E[g|\theta] = 0$, for all $\theta \in \Theta$;

(II) for almost all x , $\frac{\partial g}{\partial \theta_1}$ exists, for all $\theta \in \Theta$;

(III) $\int g \, p d\mu$ is differentiable with respect to θ_1 , and differentiation can be taken under the integral sign;

(IV) $E[\frac{\partial g}{\partial \theta_1}|\theta]$ is invertible.

The functions which satisfy conditions (I) - (IV) are called regular estimating functions with respect to θ_1 .

Let L denote the space of all regular unbiased estimating functions. For $g_1, g_2 \in L$, define the family of generalized inner products of g_1, g_2 as

$$\langle g_1, g_2 \rangle_\theta = E[g_1(X, \theta)g_2(X, \theta)^t|\theta], \quad \forall \theta \in \Theta. \quad (3.3.2)$$

Also we shall denote by s the score function of a parametric family of distributions with respect to θ_1 . We assume also that the score vector is regular in the sense described in (I) to (IV).

Definition 3.3.1. Let $(L, \langle \cdot, \cdot \rangle_\theta)$ be the family of generalized inner product spaces, and let L_0 be a subspace of L . For any $g \in L_0$, let

$$I_g(\theta) = E\left[\frac{\partial g}{\partial \theta_1}|\theta\right]^t \langle g, g \rangle_\theta^{-1} E\left[\frac{\partial g}{\partial \theta_1}|\theta\right] \quad (3.3.3)$$

An element $g^* \in L_0$ is said to be an optimal estimating function in L_0 if

$$I_{g^*}(\theta) - I_g(\theta)$$

is n. n. d., for all $g \in L_0$ and $\theta \in \Theta$.

In the rest of this section, unless otherwise stated, we shall assume the following regularity condition for estimating functions, which basically involves the interchange of differentiation and expectation.

(\mathcal{R}). For any $g \in L$,

$$E\left[\frac{\partial g}{\partial \theta}|\theta\right] = -E[g s^t|\theta]. \quad (3.3.4)$$

Thus combining (3.3.3) and (3.3.4), $I_g(\theta)$ is the information matrix of g with respect to s .

We now state the main result of this section, which is an immediate consequence of Theorem 2.2.3.

Theorem 3.3.1. Let $s = \frac{\partial \log P}{\partial \theta_1}$, $M(\theta)$ be an invertible matrix valued function, and L_0 a subspace of L . If the orthogonal projection g^* of $M(\theta) s$ into L_0 exists, then g^* is optimal in L_0 .

As an easy consequence of Theorem 3.3.1, the following result generalizes the results due to Godambe (1976), Godambe and Thompson (1974) to the multiparameter case.

Corollary 3.3.1. Suppose there exists $g : \mathcal{X} \times \Theta \rightarrow R^{d_1}$ such that

$$g^* = M(\theta) s + g \in L_0, \quad (3.3.5)$$

and g is orthogonal to every element in L_0 , then g^* is optimal in L_0 .

Proof. Since g is orthogonal to every element in L_0 , and $g^* \in L_0$, g^* is the orthogonal projection of $M(\theta) s$ into L_0 . The optimality of g^* in L_0 is immediate from Theorem 3.3.1.

Note that the above corollary generalizes the main result in Godambe and Thompson (1974) to the multiparameter case. It also provides a geometric explanation of equation (5) in Godambe and Thompson (1974). To see this, suppose there exist

matrices $\{M(\theta), M_i(\theta)\}_{i=1}^k$ of appropriate dimensions such that $M(\theta)$ is invertible and

$$g^* = M(\theta) s + \sum_{i=1}^k M_i(\theta) \frac{1}{p} \frac{\partial^i p}{\partial \theta^i} \in L_0.$$

Then g^* is the orthogonal projection of $M(\theta) s$ into L_0 .

For $d_1 = d_2 = 1$ and $k = 2$, the above result reduces to the main result in Godambe and Thompson (1974).

Next we use Theorem 3.3.1 to study the geometric ideas behind conditional and marginal inferences.

3.3.2 Geometry of Conditional Inferences

In this subsection, we study the geometry of conditional inference. As an easy consequence of this geometric approach, some of the results due to Bhapkar (1989, 1991) follow easily.

First use the identity

$$E\left[\frac{\partial g_1}{\partial \theta_1} | \theta\right] = - \langle g_1, s_{\theta_1} \rangle_{\theta},$$

where $s_{\theta_1} = \frac{\partial \log p}{\partial \theta_1}$, s_{θ_1} may involve both θ_1 and θ_2 . The information of g_1 is then

$$I_{g_1}(\theta_1; \theta) = \langle g_1, s_{\theta_1} \rangle_{\theta}^t \langle g_1, g_1 \rangle_{\theta}^{-1} \langle g_1, s_{\theta_1} \rangle_{\theta}.$$

Let L denote the space of all estimating functions which satisfy conditions (I) - (IV) of Section 3.3.1. Let L_0 be a subspace of L .

Following Bhapkar (1989, 1991a), suppose statistic (S, U) is jointly sufficient for the family $\{p_{\theta} : \theta \in \Theta\}$, and furthermore, suppose U satisfies the following condition C :

(C): The conditional distribution of S , given $u = U(X)$, depends on θ only through θ_1 , for almost all u , that is S is sufficient for the nuisance parameter θ_2 .

Denote by $h(s; \theta_1 | u)$ the conditional pdf of $S = S(X)$, given u .

Definition 3.3.2. A statistic $U = U(X)$ is said to be partially ancillary for θ_1 in the complete sense if

(i) U satisfies requirement C;

(ii) the family $\{p_\theta^U : \theta_2 \in \Theta_2\}$ of distributions of U for fixed θ_1 is complete for every $\theta_1 \in \Theta_1$.

A statistic $U = U(X)$ is said to be partially ancillary for θ_1 in the weak sense if

(i) U satisfies requirement C;

(ii) the marginal distribution of U depends on θ only through a parametric function $\delta = \delta(\theta)$ (δ is assumed to be differentiable) such that (θ_1, δ) is a one-to-one function of θ .

Letting p_θ^U be the pdf of U and

$$l_c(x; \theta_1) = \frac{\partial \log h(s, \theta_1 | u)}{\partial \theta_1},$$

the following theorem connects Theorem 3.3.1 and Bhapkar's (1989, 1991a) results. Take $L = L_0$.

Theorem 3.3.2. If the statistic $U = U(X)$ is partially ancillary for θ_1 in the complete sense or in the weak sense, then $l_c(x; \theta_1)$ is the orthogonal projection of s_{θ_1} into L .

Proof. We are going to prove this result in two cases.

Case 1. U is partially ancillary for θ_1 in the complete sense.

Note that

$$s_{\theta_1} = l_c(x; \theta_1) + \frac{\partial \log p_\theta^U}{\partial \theta_1}.$$

We only need to show that $\forall g_1 \in L, \theta \in \Theta$

$$\langle g_1, \frac{\partial \log p_\theta^U}{\partial \theta_1} \rangle_\theta = 0.$$

But

$$\begin{aligned} \langle g_1, \frac{\partial \log p_\theta^U}{\partial \theta_1} \rangle_\theta &= E[g_1 \left(\frac{\partial \log p_\theta^U}{\partial \theta_1} \right)^t | \theta] \\ &= E\{E[g_1 \left(\frac{\partial \log p_\theta^U}{\partial \theta_1} \right)^t | \theta, U] | \theta\} \\ &= E\{E[g_1 | \theta, U] \left(\frac{\partial \log p_\theta^U}{\partial \theta_1} \right)^t | \theta\}. \end{aligned}$$

Since $E\{E[g_1 | \theta, U] | \theta\} = 0$, by the completeness of $\{p_\theta^U : \theta_2 \in \Theta_2\}$ for fixed $\theta_1 \in \Theta_1$, $E[g_1 | \theta, U] = 0$ almost everywhere for fixed θ_1 . This implies that

$$\langle g_1, \frac{\partial \log p_\theta^U}{\partial \theta_1} \rangle_\theta = 0, \quad \forall \theta \in \Theta.$$

Thus $l_c(x; \theta_1)$ is the orthogonal projection of s_{θ_1} into L .

Case 2. U is partially ancillary for θ_1 in the weak sense.

Again note that

$$s_{\theta_1} = l_c(x; \theta_1) + \frac{\partial \log p_\theta^U}{\partial \theta_1}.$$

Now since the map $(\theta_1, \theta_2) \rightarrow (\theta_1, \delta(\theta))$ is a one-to-one map, the matrix

$$\begin{bmatrix} I_{d_1} & \frac{\partial \delta}{\partial \theta_1} \\ 0 & \frac{\partial \delta}{\partial \theta_2} \end{bmatrix}.$$

is invertible. It implies that $\frac{\partial \delta}{\partial \theta_2}$ is invertible. Note that

$$\frac{\partial \log p_\theta^U}{\partial \theta_1} = \frac{\partial \log p_\theta^U}{\partial \theta_2} \left(\frac{\partial \delta}{\partial \theta_2} \right)^{-1} \frac{\partial \delta}{\partial \theta_1}$$

$$= \frac{\partial \log p_{\theta}^U}{\partial \theta_2} D(\theta),$$

where $D(\theta) = (\frac{\partial \delta}{\partial \theta_2})^{-1} \frac{\partial \delta}{\partial \theta_1}$. Thus we only need to show that $\forall g_1 \in L, \theta \in \Theta$,

$$\langle g_1, \frac{\partial \log p_{\theta}^U}{\partial \theta_2} \rangle_{\theta} = 0.$$

But this follows easily by differentiating

$$E[g_1|\theta] = 0$$

with respect to θ_2 , and using the regularity conditions.

By combining Theorems 3.3.1 and 3.3.2, we get the following result due to Bhapkar (1989, 1991a).

Corollary 3.3.2. With the notation as above, $l_c(x; \theta_1)$ is optimal in L if either U is partially ancillary for θ_1 in the complete sense or in the weak sense, that is

$$I_g(\theta) \leq I_{l_c}(\theta), \quad \forall \theta \in \Theta, g \in L.$$

Note that from the proof of Theorem 3.3.2, the condition of partial ancillarity for θ_1 in either the complete sense or in weak sense guarantees that the conditional score is the orthogonal projection of the score function with respect to θ_1 into L .

3.3.3 Geometry of Marginal Inference

.

In this subsection, we study the geometry of marginal inference. As an easy consequence of our approach, the optimality result of Bhapkar (1989) and Lloyd (1987) on marginal inference will follow easily. Assume that

(M): the distribution of statistic $S = S(X)$ depends on θ only through θ_1 .

Definition 3.3.3. A statistic $S = S(X)$ is said to be partially sufficient for θ_1 in the complete sense if

(i) S satisfies condition (M);

(ii) given $s = S(X)$, the family $\{p_\theta^{U|s} : \theta_2 \in \Theta_2\}$ of the conditional distributions of U for fixed θ_1 is complete for almost all s , and for every $\theta_1 \in \Theta_1$.

A statistic $S = S(X)$ is said to be partially sufficient for θ_1 in the weak sense if

(i) S satisfies condition (M);

(ii) the conditional distribution of U , given $s = S(X)$, depends on θ only through a parametric function $\delta = \delta(\theta)$ (δ is assumed to be differentiable) such that (θ_1, δ) is a one-to-one function of θ .

If $S = S(X)$ is partially sufficient for θ_1 in the complete (or weak) sense. Let $p_\theta^{U|S}$ denote the conditional pdf of U given $S = s$, and

$$l_m(x; \theta_1) = \frac{\partial \log f(s; \theta_1)}{\partial \theta_1}.$$

The main result of this subsection is given in the following theorem.

Theorem 3.3.3. If the statistic $S = S(X)$ is partially sufficient for θ_1 in the complete sense or in the weak sense, then $l_m(x; \theta_1)$ is the orthogonal projection of s_{θ_1} into L .

Proof. We are going to prove this result in two cases.

Case 1. S is partially sufficient for θ_1 in the complete sense.

Note that

$$s_{\theta_1} = l_m(x; \theta_1) + \frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_1}.$$

We only need to show that for any $g_1 \in L, \theta \in \Theta$

$$\langle g_1, \frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_1} \rangle_\theta = 0.$$

But

$$\langle g_1, \frac{\partial \log p_\theta^{U|s}}{\partial \theta_1} \rangle_\theta = E[g_1 \left(\frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_1} \right)^t | \theta]$$

$$\begin{aligned}
&= E\{E[g_1] \left(\frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_1}\right)^t | \theta, S | \theta\} \\
&= E\{E[g_1 | \theta, S] \left(\frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_1}\right)^t | \theta\}.
\end{aligned}$$

Since $E\{E[g_1 | \theta, S] | \theta\} = 0$, by the completeness of $\{p_\theta^{(U|s)} : \theta_2 \in \Theta_2\}$ for fixed $\theta_1 \in \Theta_1$, $E[g_1 | \theta, S] = 0$ for fixed θ_1 . This implies that

$$\langle g_1, \frac{\partial \log p_\theta^U}{\partial \theta_1} \rangle_\theta = 0, \quad \forall \theta \in \Theta.$$

Thus $l_m^t(x; \theta_1)$ is the orthogonal projection of s_{θ_1} into L .

Case 2. S is partially sufficient for θ_1 in the weak sense.

Again note that

$$s_{\theta_1} = l_m(x; \theta_1) + \frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_1}.$$

Now since the map $(\theta_1, \theta_2) \longrightarrow (\theta_1, \delta(\theta))$ is a one-to-one map, the matrix

$$\begin{bmatrix} I_{d_1} & \frac{\partial \delta}{\partial \theta_1} \\ 0 & \frac{\partial \delta}{\partial \theta_2} \end{bmatrix}.$$

is invertible. This implies that $\frac{\partial \delta}{\partial \theta_2}$ is invertible. Hence

$$\begin{aligned}
\frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_1} &= \frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_2} \left(\frac{\partial \delta}{\partial \theta_2}\right)^{-1} \frac{\partial \delta}{\partial \theta_1} \\
&= \frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_2} N(\theta),
\end{aligned}$$

where $N(\theta) = \left(\frac{\partial \delta}{\partial \theta_2}\right)^{-1} \frac{\partial \delta}{\partial \theta_1}$. Thus we only need to show that $\forall g_1 \in L, \theta \in \Theta$,

$$\langle g_1, \frac{\partial \log p_\theta^{(U|s)}}{\partial \theta_2} \rangle_\theta = 0.$$

But this follows easily by differentiating

$$E[g_1|\theta] = 0$$

with respect to θ_2 , and using the regularity conditions. Thus $l_m(x; \theta_1)$ is the orthogonal projection of s_{θ_1} into L . This completes the proof.

By combining Theorem 3.3.1 and 3.3.3, we get $I_g(\theta) \leq I_{l_m}(\theta)$ for all $\theta \in \Theta, g \in L$, which includes the results of Bhapkar(1989, 1991a) and Lloyd (1987) as a corollary.

Corollary 3.3.3.

(1) if S is partially ancillary for θ_1 in the complete sense, then $l_m(x; \theta_1)$ is optimal in L ;

(2) if S is partially ancillary for θ_1 in the weak sense, then $l_m(x; \theta_1)$ is optimal in L .

Proof. In both cases, $l_m(x; \theta_1)$ is the orthogonal projection of s_{θ_1} into L . So the inequality

$$I_g(\theta) \leq I_{l_m}(\theta), \quad \forall \theta \in \Theta, g \in L,$$

follows from Theorem 3.3.1.

Note that (1) of the above corollary is the main result of Lloyd (1987) and (2) is due to Bhapkar (1989, 1991a).

3.4 Locally Optimal Estimating Functions

In this section, a general result about locally optimality of estimating functions in the presence of nuisance parameters will be proved. As easy consequences of this result, some results of Lindsay (1982), Godambe (1985), Murphy and Li (1995) will be studied.

3.4.1 A General Result

In this subsection, we study the local optimal estimating functions in the presence of nuisance parameters. We first introduce the space of estimating functions of interest.

Let \mathcal{X} be a sample space, $\Theta = \Theta_1 \times \Theta_2$ the $d_1 + d_2$ dimensional parameter space, with $\Theta_i \subset R^{d_i}$ ($i = 1, 2$). A function

$$g : \mathcal{X} \times \Theta \longrightarrow R^{d_1}$$

is said to be an unbiased estimating function if

$$E[g(X, \theta) | \theta] = 0, \quad \forall \theta = (\theta_1, \theta_2) \in \Theta.$$

An estimating function g is said to be regular if it satisfies conditions (I) - (IV) as in Section 3.3.1.

Let L denote the space of all regular unbiased estimating functions from $\mathcal{X} \times \Theta$ to R^{d_1} . Also the score vector s_{θ_1} is assumed to be regular in the sense described in (i) and (ii).

Definition 3.4.1. Let $(L, < \cdot, \cdot >_\theta)$ be the family of generalized inner product spaces, and let L_0 be a subspace of L . For any $g \in L_0$, the information function of g is defined by

$$I_g(\theta) = E\left[\frac{\partial g}{\partial \theta_1} | \theta\right]^t < g, g >_\theta^{-1} E\left[\frac{\partial g}{\partial \theta_1} | \theta\right] \quad (3.4.6)$$

An element $g^* \in L_0$ is said to be a locally optimal estimating function at $\theta_2 = \theta_{20}$ if

$$I_{g^*}(\theta_1, \theta_{20}) - I_g(\theta_1, \theta_{20})$$

is n. n. d., for all $g \in L_0$ and $\theta_1 \in \Theta_1$.

The following is the main result of this section.

Theorem 3.4.1. Let L be the space of all regular unbiased estimating functions

$$g : \mathcal{X} \times \Theta \longrightarrow R^{d_1}.$$

Let L_0 be subspace of L . If g^* is the orthogonal projection of s into L_0 with respect to the generalized inner products $\langle \cdot, \cdot \rangle_\theta$ with $\theta_2 = \theta_{20}$, then g^* is locally optimal in L_0 , that is for any fixed $\theta_{20} \in \Theta_2$,

$$I_{g^*}(\theta_1, \theta_{20}) \geq I_g(\theta_1, \theta_{20}),$$

for all $\theta_1 \in \Theta_1$. Also the local optimal estimating function in L_0 is unique in the following sense: if $g \in L_0$, then $I_{g^*}(\theta_1, \theta_{20}) = I_g(\theta_1, \theta_{20})$ for all $\theta_1 \in \Theta_1$ if and only if there exists an invertible matrix valued function $N : \Theta_1 \times \{\theta_{20}\} \longrightarrow M_{d_1 \times d_1}$ such that for all $\theta_1 \in \Theta_1$,

$$g^*(X; \theta_1, \theta_{20}) = N(\theta_1, \theta_{20}) g(X; \theta_1, \theta_{20}),$$

with probability 1 with respect to $P_{\theta_1, \theta_{20}}$.

Proof. This follows easily from Theorem 2.2.4 and 3.3.1.

Next we apply Theorem 3.4.1 to generalize the results in three different cases:

- (1) Lindsay's (1982) result on the local optimality of conditional score functions;
- (2) Godambe's (1985) result on the estimation in stochastic processes;
- (3) Murphy and Li's (1995) result on the projected partial likelihood.

3.4.2 Local Optimality of Conditional Score Functions

.

Suppose that $(X_1, \dots, X_n) = X_{(n)}$ is a sequence of possibly dependent observations with pdf

$$f(X_{(n)}; \theta) = f_1(X_{(1)}; \theta) f_2(X_{(2)} | X_{(1)}; \theta) \dots f_n(X_{(n)} | X_{(n-1)}; \theta). \quad (3.4.7)$$

Let $U_i = \frac{\partial}{\partial \theta_1} \log f_j$, and $S_j(\theta_1) = S_j(X_{(j)}; \theta_1)$ be minimal sufficient for θ_2 with θ_1 fixed in pdf f_j , let

$$W_j = U_j - E[U_j | S_j, X_{(j-1)}], \quad (3.4.8)$$

for $j \in \{1, \dots, n\}$. The sequence $\{S_j(\theta_1)\}_{j=1}^n$ is called sequentially complete if for each k from 1 to n , the system of equalities

$$E[H(S_k, X_{(k-1)}; \theta_1)] = 0 \quad (3.4.9)$$

for all θ with θ_1 fixed implies that $H(S_k, X_{(k-1)}; \theta_1)$ is a constant in S_k with probability one, that is, $H(S_k, X_{(k-1)}; \theta_1)$ does not depend on S_k . The following is a slight generalization of the main result in Lindsay (1982).

Theorem 3.4.2. Assume sequential completeness and $E[U_i | X_{(i-1)}] = 0$ for all $i = 1, \dots, n$. For fixed $\theta_{20} \in \Theta_2$, consider unbiased estimating function

$$h : \mathcal{X} \times \Theta_1 \times \{\theta_{20}\} \longrightarrow R^{d_1}, \quad (3.4.10)$$

Let L_0 be the subspace, which consists of all unbiased estimating functions from $\mathcal{X} \times \Theta_1 \times \{\theta_{20}\}$ into R^{d_1} . Then

(a) $W(\theta_1, \theta_{20}) = \sum_{i=1}^n W_i$ is the orthogonal projection of s_{θ_1} into L_0 with respect to the generalized inner product $\langle \cdot, \cdot \rangle_{\theta_1, \theta_{20}}$;

(b) $W(\theta_1, \theta_{20})$ is optimal in L_0 , and the optimal element in L_0 is unique in the following sense: if $g \in L_0$ and $I_g(\theta_1, \theta_{20}) = I_W(\theta_1, \theta_{20})$ for all $\theta_1 \in \Theta_1$, then there exists an invertible matrix valued function $N : \Theta_1 \times \{\theta_{20}\} \longrightarrow M_{d_1 \times d_1}$ such that for all $\theta_1 \in \Theta_1$

$$W = N(\theta_1, \theta_{20}) g,$$

with probability one with respect to $P_{\theta_1, \theta_{20}}$.

Proof. First note that, for any $H \in L_0$, consider the decomposition

$$H = H_n + H_{n-1} + \dots + H_1 + H_0,$$

where

$$H_n(X_{(n)}; \theta_1) = H - E[H|S_n, X_{(n-1)}],$$

$$H_{n-1}(S_n, X_{(n-1)}; \theta_1) = E[H|S_n, X_{(n-1)}] - E[H|S_{n-1}, X_{(n-2)}],$$

$$H_k(S_{k+1}, X_{(k)}; \theta_1) = E[H|S_{k+1}, X_{(k)}] - E[H|S_k, X_{(k-1)}],$$

for all $k \in \{1, 2, \dots, n\}$, and $H_0 = E[H|S_1]$. By the sequential completeness of $\{S_j(\theta_1)\}_{j=1}^{j=n}$, $H_k(S_{k+1}, X_{(k)}; \theta_1)$ does not depend on S_{k+1} , that is $H_k(S_{k+1}, X_{(k)}; \theta_1) = H_k(X_{(k)}; \theta_1)$.

(a) Since

$$s_{\theta_1} = \sum_{j=1}^n U_j = W + \sum_{j=1}^n E[U_j|S_j, X_{j-1}],$$

it suffices to show that

$$< E[U_j|S_j, X_{j-1}], H_k(S_{k+1}, X_{(k)}; \theta_1) >_{\theta} = 0,$$

for all $j, k \in \{1, \dots, n\}$. But this follows from an easy conditioning argument.

(b) This follows from part (a) and Theorem 3.4.1.

Note that part (a) of Theorem 3.4.2 is a restatement of the main result in Lindsay (1982).

3.4.3 Locally Optimal Estimating Functions for Stochastic Processes

In this subsection, we generalize the results of Godambe (1985) to the case where there are nuisance parameters. As a special case, we get generalized estimating equations in the presence of nuisance parameters.

Let $\{X_1, X_2, \dots, X_n\}$ be a discrete stochastic process, $\Theta_j \subset R^{d_j}$ ($j = 1, 2$) be open sets. Let h_i be a R^{d_1} valued function of X_1, \dots, X_i and θ , which satisfies for fixed $\theta_{20} \in \Theta_2$,

$$E_{i-1}[h_i(X_1, \dots, X_i; \theta)|\theta_1, \theta_{20}] = 0, \quad (i = 1, \dots, n, \theta_1 \in \Theta_1). \quad (3.4.11)$$

In the above, E_{i-1} denotes the conditional expectation conditioning on the first $i - 1$ variables, namely, X_1, \dots, X_{i-1} . Let

$$L_0 = \{g : g = \sum_{i=1}^n A_{i-1} h_i\},$$

where A_{i-1} is a $M_{k \times k}$ valued function of X_1, \dots, X_{i-1} and θ_1 , for all $i \in \{1, \dots, n\}$.

The following theorem, which generalizes the result of Godambe (1985) on optimal estimating functions for stochastic processes.

Theorem 3.4.3. Let $\theta_2 = \theta_{20}$, and suppose h_i satisfies the regularity condition (\mathcal{R}). Let

$$A_i^* = E_{i-1} \left[\frac{\partial h_i}{\partial \theta_1} | \theta_1, \theta_{20} \right]^t E_{i-1} [h_i h_i^t | \theta_1, \theta_{20}]^{-1} \quad \forall i \in \{1, 2, \dots, n\},$$

and

$$g^* = \sum_{i=1}^n A_i^* h_i,$$

then the following conclusions hold:

(a). g^* is the orthogonal projection of s_{θ_1} into L_0 with respect to the generalized inner product $\langle \cdot, \cdot \rangle_{\theta_1, \theta_{20}}$.

(b). g^* is a locally optimal estimating function in L_0 , i. e.,

$$I_g(\theta_1, \theta_{20}) \leq I_{g^*}(\theta_1, \theta_{20}),$$

for all $g \in L_0$ and $\theta_1 \in \Theta_1$.

(c). If $g \in L_0$ and $E[g g^t | \theta]$ is invertible, then $I_g(\theta_1, \theta_{20}) = I_{g^*}(\theta_1, \theta_{20})$, $\forall \theta_1 \in \Theta_1$ if and only if there exists an invertible matrix function $N : \Theta_1 \times \{\theta_{20}\} \rightarrow M_{k \times k}$ such that for any $\theta_1 \in \Theta_1$,

$$g_*(X_1, \dots, X_n; \theta_1, \theta_{20}) = N(\theta_1, \theta_{20}) g(X_1, \dots, X_n; \theta_1, \theta_{20}),$$

with probability 1 with respect to $P_{\theta_1, \theta_{20}}$.

Proof. (a). For any $g = \sum_{i=1}^n A_i$ $h_i \in L_0$, $\theta_1 \in \Theta_1$,

$$\begin{aligned} & \langle s_{\theta_1} - g^*, g \rangle_{(\theta_1, \theta_{20})} = \langle s_{\theta_1}, g \rangle_{(\theta_1, \theta_{20})} - \langle g^*, g \rangle_{(\theta_1, \theta_{20})} \\ &= \sum_{i=1}^n E[s_{\theta_1} h_i^t A_i^t | (\theta_1, \theta_{20})] - \sum_{i=1}^n \sum_{j=1}^n E[A_i^* h_i h_j^t A_j^t | (\theta_1, \theta_{20})] \\ &= \sum_{i=1}^n E\{E_{i-1}[s_{\theta_1} h_i^t A_i^t | (\theta_1, \theta_{20})] | (\theta_1, \theta_{20})\} - \sum_{i=1}^n E[A_i^* h_i h_i^t A_i^t | (\theta_1, \theta_{20})] \\ &\quad - \sum_{i < j} E[A_i^* h_i h_j^t A_j^t | (\theta_1, \theta_{20})] - \sum_{i > j} E[A_i^* h_i h_j^t A_j^t | (\theta_1, \theta_{20})]. \end{aligned} \quad (3.4.12)$$

But for $i < j$,

$$\begin{aligned} E[A_i^* h_i h_j^t A_j^t | (\theta_1, \theta_{20})] &= E\{E_{j-1}[A_i^* h_i h_j^t A_j^t | (\theta_1, \theta_{20})] | (\theta_1, \theta_{20})\} \\ &= E\{A_i^* h_i E_{j-1}[h_j^t A_j^t | (\theta_1, \theta_{20})] | (\theta_1, \theta_{20})\} = 0. \end{aligned}$$

Similarly, for $i > j$,

$$E[A_i^* h_i h_j^t A_j^t | (\theta_1, \theta_{20})] = 0.$$

Thus from equation (3.4.12), we get

$$\begin{aligned} \langle s_{\theta_1} - g^*, g \rangle_{\theta_1, \theta_{20}} &= \sum_{i=1}^n E\{E_{i-1}[\frac{\partial h_i}{\partial \theta_1} | \theta_1, \theta_{20}]^t A_i^t | \theta_1, \theta_{20}\} - \\ &\quad \sum_{i=1}^n E\{A_i^* E_{i-1}[h_i h_i^t | \theta_1, \theta_{20}] A_i^t | \theta_1, \theta_{20}\} = 0. \end{aligned}$$

Hence g^* is the orthogonal projection of s into L_0 .

Parts (b) and (c) of the theorem follows easily from part (a) and Theorem 3.4.1.

As a corollary to Theorem 3.4.3, the generalized estimating equations in the presence of nuisance parameters for multivariate data can be easily obtained.

Corollary 3.4.1. Suppose that X_1, \dots, X_n are independent, for fixed $\theta_{20} \in \Theta_2$, for each $i \in \{1, 2, \dots, n\}$,

$$h_i : \mathcal{X}_i \times \Theta \longrightarrow R^{d_1},$$

with $E[h_i(X_i, \theta) | \theta_1, \theta_{20}] = 0$. Consider the subspace L_0 as that in Theorem 3.4.3.

Then the generalized estimating equations determined by $\{h_i\}$ is given by

$$\sum_{i=1}^n A_i^* h_i = 0,$$

where

$$A_i^* = E_{i-1} \left[\frac{\partial h_i}{\partial \theta_1} | \theta_1, \theta_{20} \right]^t E_{i-1} [h_i h_i^t | \theta_1, \theta_{20}]^{-1} \quad \forall i \in \{1, 2, \dots, n\}.$$

The above corollary provides a very convenient way to construct generalized estimating equations. For instance, if h_i is chosen as linear (or quadratic) function of X_i , then the corresponding generalized estimating equations reduce to the GEE1 and GEE2 studied by Liang, Zeger and their associates. One may refer to see Liang and Zeger (1986), Liang, Zeger and Qaqish (1992), Diggle, Liang and Zeger (1994).

3.4.4 Local Optimality of Projected Partial Likelihood

In this subsection, we generalize the result of Murphy and Li (1995) on projected partial likelihood to the nuisance parameter case. Also the application of this result to longitudinal data will be pointed out.

Suppose that the data consist of a vector of observations X with density $f(x; \theta_1, \theta_2)$, θ_1 is the vector of parameters of interest, which is finite dimensional, and θ_2 is the vector of nuisance parameters, which may be infinite dimensional. Suppose there is a one-to-one transformation of the data X into a set of variables $Y_1, C_1, \dots, Y_m, C_m$. Let

$$Y^{(j)} = (Y_1, \dots, Y_j), \quad C^{(j)} = (C_1, \dots, C_j), \quad j = 1, \dots, m. \quad (3.4.13)$$

For instance, in survival analysis, Y_1, \dots, Y_m denote the lifetime variables, and C_1, \dots, C_m the censoring variables.

Note that the joint density of $Y^{(m)}, C^{(m)}$ can be written as

$$\prod_{j=1}^m f(c_j | c^{(j-1)}, y^{(j-1)}; \theta_1, \theta_2) \prod_{j=1}^m f(y_j | c^{(j)}, y^{(j-1)}; \theta_1, \theta_2), \quad (3.4.14)$$

where $c^{(0)}$ and $y^{(0)}$ are arbitrary constants, and are used only for notational purposes.

Then $P(\theta_1) = \prod_{j=1}^m f(y_j | c^{(j)}, y^{(j-1)}; \theta_1, \theta_2)$ is called the Cox partial likelihood.

Let

$$s_{\theta_1} = \sum_{j=1}^m \frac{\partial \log f(c_j | c^{(j-1)}, y^{(j-1)}; \theta_1, \theta_2)}{\partial \theta_1} + s^*,$$

where

$$s^* = \sum_{j=1}^m \frac{\partial \log f(y_j | c^{(j)}, y^{(j-1)}; \theta_1, \theta_2)}{\partial \theta_1}. \quad (3.4.15)$$

Next we introduce the subspace of unbiased estimating functions which is of interest in this subsection. This is similar to the space considered by Godambe (1985) in studying the foundation of finite sample estimation in stochastic processes.

For any $j \in \{1, 2, \dots, m\}$, consider estimating functions

$$h_j : Y^{(j)} \times C^{(j)} \times \Theta_1 \longrightarrow R^{d_1}, \quad (3.4.16)$$

$$E[h_j | y^{(j-1)}, c^{(j)}, \theta] = 0, \quad (3.4.17)$$

for all $\theta \in \Theta_1 \times \Theta_2$.

For chosen $\{h_i\}_{i=1}^m$, consider the space

$$L_0 = \{g : g = \sum_{j=1}^m A_j(\theta) h_j\}, \quad (3.4.18)$$

where for all $j \in \{1, \dots, m\}$, $A_j(\theta)$ is a $d_1 \times d_1$ matrix, and h_j satisfies (3.4.16) and (3.4.17).

Let

$$s^* = \frac{\partial \log P}{\partial \theta_1} = \sum_{j=1}^m \frac{\partial \log f(y_j | c^{(j)}, y^{(j-1)}; \theta_1, \theta_2)}{\partial \theta_1}. \quad (3.4.19)$$

The main result of this subsection is the following.

Theorem 3.4.4. For fixed $\theta_2 = \theta_{20}$, for any $i \in \{1, \dots, m\}$, let

$$A_i^* = E\left[\frac{\partial h_i}{\partial \theta_1} | \theta_1, \theta_{20}, y^{(i-1)}, c^{(i)}\right]^t E[h_i h_i^t | \theta_1, \theta_{20}, y^{(i-1)}, c^{(i)}]^{-1},$$

and

$$g^* = \sum_{i=1}^m A_i^* h_i.$$

Then

(a) g^* is the orthogonal projection of s^* into L_0 , that is, for any $g \in L_0$,

$$\langle s^* - g^*, g \rangle_{\theta_1, \theta_{20}} = 0,$$

for all $\theta \in \Theta_1$;

(b) g^* is locally optimal in L_0 at $\theta_2 = \theta_{20}$;

(c) If $g \in L_0$ and $E[g g^t | \theta]$ is invertible, then $I_g(\theta_1, \theta_{20}) = I_{g^*}(\theta_1, \theta_{20})$, $\forall \theta_1 \in \Theta_1$ if and only if there exists an invertible matrix function $N : \Theta_1 \times \{\theta_{20}\} \rightarrow M_{k \times k}$ such that for any $\theta_1 \in \Theta_1$,

$$g_*(X; \theta_1, \theta_{20}) = N(\theta_1, \theta_{20}) g(X; \theta_1, \theta_{20}),$$

with probability 1 with respect to $P_{\theta_1, \theta_{20}}$.

Proof. For any $j \in \{1, \dots, m\}$, let

$$s_j = \frac{\partial \log f(c_j | c^{(j-1)}, y^{(j-1)}, \theta_1, \theta_2)}{\partial \theta_1},$$

then $s - s^* = \sum_{j=1}^m s_j$.

(a) For any $g \in L_0$, let $g_j = A_j(\theta) h_j, j = 1, \dots, m$ be the components in the definition of L_0 . In order to show that $E[(s - s^*) g^t | \theta] = 0$, it suffices to prove that

$$E[s_j g_{j'}^t | \theta] = 0,$$

for all $j, j' \in \{1, \dots, m\}$. Consider the following three cases:

Case 1. $j > j'$. Then

$$E[s_j g_{j'}^t | \theta] = E\{E[s_j | c^{(j-1)}, y^{(j-1)}] g_{j'}^t | \theta\} = 0,$$

since $E[s_j | c^{(j-1)}, y^{(j-1)}] = 0$.

Case 2. $j = j'$. Then

$$E[s_j g_j^t | \theta] = E\{s_j E[g_j | y^{(j-1)}, c^{(j)}]^t | \theta\} = 0,$$

since $E[g_j | y^{(j-1)}, c^{(j)}]^t = 0$.

Case 3. $j' > j$. Then

$$E[s_j g_{j'}^t | \theta] = E\{s_j E[g_{j'} | y^{(j'-1)}, c^{(j'-1)}]^t | \theta\} = 0,$$

since $E[g_{j'} | y^{(j'-1)}, c^{(j'-1)}] = 0$.

Part (b) and (c) follows from Theorem 3.4.1 and part (a).

Note that Murphy and Li (1995) studied projected partial likelihood in the case $d_1 = 1$, the absence of nuisance parameters and when all the $C^{(j)}$'s are empty. Similar to Murphy and Li's comments, because of the nested structure of $\{Y^{(j)}, C^{(j)}\}$, removing drop-out factors in this way will not cause bias in the resulting partial score function, as long as the subjects' drop-out depends only the past. This is in contrast to a generalized estimating equation, which is biased under random drop-out.

3.5 Optimal Conditional Estimating Functions

In this section, we study optimal conditional estimating functions. Let \mathcal{X} be a sample space, $\Theta = \Theta_1 \times \Theta_2$, $\Theta_i \subset R^{d_i}$, ($i = 1, 2$), and θ_1 is the parameter of interest. For fixed θ_1 , assume that $S(\theta_1)$ is sufficient for θ_2 . A function

$$g : \mathcal{X} \times \Theta_1 \longrightarrow R^{d_1},$$

is called a regular conditional unbiased estimating function if

- (1) $E[g | S(\theta_1)] = 0$, for all $\theta_1 \in \Theta_1$;
- (2) $E[g g^t | S(\theta_1)]$ is positive definite.

Consider the space L of all regular conditional unbiased estimating functions, a family of generalized inner products on L is defined as follows: for any $g_1, g_2 \in L$,

$$\langle g_1, g_2 \rangle_{S(\theta_1)} = E[g_1 g_2^t | S(\theta_1)]. \quad (3.5.20)$$

For any $g \in L$, the conditional information of g is defined as follows:

$$I_g(\theta_1|S(\theta_1)) = E\left[\frac{\partial g}{\partial \theta_1}|S(\theta_1)]^t < g, g >_{S(\theta_1)}^{-1} E\left[\frac{\partial g}{\partial \theta_1}|S(\theta_1)]\right]. \quad (3.5.21)$$

Definition 3.5.1. Let L_0 be a subspace of L , a function $g^* \in L_0$ is called an optimal conditional estimating function in L_0 if

$$I_{g^*}(\theta_1|S(\theta_1)) \geq I_g(\theta_1|S(\theta_1)),$$

for all $g \in L_0$.

The main result in this section is given in the following theorem.

Theorem 3.5.1. Define

$$s = \frac{\partial}{\partial \theta_1} \log f(X|S(\theta_1)). \quad (3.5.22)$$

Suppose L_0 is a subspace of L , and assume that the orthogonal projection g^* of s into L_0 exists. Then

(a) g^* is an optimal conditional estimating function in L_0 ;

(b) the optimal element in L_0 is unique in the sense that if $g \in L_0$, then $I_{g^*}(\theta_1|S(\theta_1)) = I_g(\theta_1|S(\theta_1))$, for all $\theta_1 \in \Theta_1$ if and only if there exists an invertible matrix valued function $N : \mathcal{X} \times \Theta_1 \longrightarrow M_{k \times k}$ of the form $N(X, \theta_1) = N(S(\theta_1))$ such that

$$g^*(X; \theta_1) = N(S(\theta_1)) g(X; \theta_1).$$

Proof. (a) Since $E[g|S(\theta_1)] = 0$, under the regularity condition given in (2.5),

$$E\left[\frac{\partial g}{\partial \theta_1}|S(\theta_1)] = -E[g s^t|S(\theta_1)].$$

Thus, from the definition of $I_g(\theta_1|S(\theta_1))$,

$$I_g(\theta_1|S(\theta_1)) = < g, s >_{S(\theta_1)}^t < g, g >_{S(\theta_1)}^{-1} < g, s >_{S(\theta_1)}.$$

Since g^* is the orthogonal projection of s into L_0 with respect to $\langle \cdot, \cdot \rangle_{S(\theta_1)}$, thus the optimality of g^* in L_0 follows from Theorem 2.2.3.

(b) This follows from Theorem 2.2.4.

Next as applications of Theorem 3.5.1, we generalize the results of Godambe and Thompson (1989) on optimal estimating functions into the conditional estimating functions framework.

To this end, let \mathcal{X} denote the sample space, $\theta_1 = (\theta_{11}, \dots, \theta_{1d_1})$ be a vector of parameters, $h_j, j = 1, \dots, k$ be real functions on $\mathcal{X} \times \Theta_1$ such that

$$E[h_j(X, \theta) | S(\theta_1), \mathcal{X}_j] = 0, \quad \forall \theta \in \Theta, \quad j = 1, \dots, k,$$

where \mathcal{X}_j be a specified partition of $\mathcal{X}, j = 1, \dots, k$. We will denote

$$E[\cdot | S(\theta_1), \mathcal{X}_j] = E_{(j)}[\cdot | S(\theta_1)].$$

Consider the class of estimating functions

$$L_0 = \{g : g = (g_1, \dots, g_{d_1})\}$$

where

$$g_r = \sum_{j=1}^k q_{jr} h_j, \quad r = 1, \dots, m,$$

$q_{jr} : \mathcal{X} \times \Theta_1 \rightarrow R$ being measurable with respect to the partition \mathcal{X}_j for $j = 1, \dots, k, r = 1, \dots, d_1$.

Let

$$q_{jr}^* = \frac{E_{(j)}[\frac{\partial h_j}{\partial \theta_r} | S(\theta_1)]}{E_{(j)}[h_j^2 | S(\theta_1)]}, \quad (3.5.23)$$

for all $j = 1, \dots, k, r = 1, \dots, d_1$, and

$$g_r^* = \sum_{j=1}^k q_{jr}^* h_j, \quad r = 1, \dots, d_1.$$

The estimating functions $h_j, j = 1, \dots, k$ are said to be *mutually orthogonal* if

$$E_{(j)}[q_{jr}^* h_j q_{j'r'}^* h_{j'} | S(\theta_1)] = 0, \quad \forall j \neq j', r, r' = 1, \dots, d_1. \quad (3.5.24)$$

Theorem 3.5.2. Suppose $\{h_j\}_{j=1}^k$ are mutually orthogonal. Then the following hold:

(a) g^* is the orthogonal projection of the score function s into L_0 .

(b) g^* is an optimal estimating function in L_0 .

(c). If $g \in L_0$, and $E[g g^t | \theta]$ is invertible, then $I_g(\theta) = I_{g^*}(\theta)$, $\forall \theta \in \Theta$ if and only if there exists an invertible matrix function $N : \mathcal{X} \times \Theta_1 \rightarrow M_{k \times k}$ of the form $N(X, \theta_1) = N(S(\theta_1))$ such that for any $\theta_1 \in \Theta_1$,

$$g^*(X; \theta_1) = N(S(\theta_1)) g(X; \theta_1),$$

with probability 1 with respect to P_θ .

Proof. (1). We only need to show that, $\forall r \in \{1, \dots, d_1\}$, $g_r = \sum_{j=1}^k q_{jr} h_j$,

$$\langle s - g_r^*, g_r \rangle_{S(\theta_1)} = 0, \quad \forall \theta_1 \in \Theta_1$$

that is

$$\langle s, g_r \rangle_{S(\theta_1)} = \langle g_r^*, g_r \rangle_{S(\theta_1)}, \quad \forall \theta_1 \in \Theta_1.$$

But

$$\begin{aligned} \langle g_r^*, g_r \rangle_{S(\theta_1)} &= \sum_{j=1}^k \sum_{j'=1}^k E[q_{jr}^* h_j q_{j'r} h_{j'} | S(\theta_1)] \\ &= \sum_{j=1}^k \sum_{j'=1}^k E\{q_{j'r}^{-1} h_j q_{j'r} E_{(j)}[q_{jr}^* h_j q_{j'r}^* h_{j'} | S(\theta_1)] | S(\theta_1)\} \\ &= \sum_{j=1}^k E\{q_{jr}^* q_{jr} E_{(j)}[h_j^2 | S(\theta_1)] | S(\theta_1)\} \\ &= \sum_{j=1}^k E\{q_{jr} E_{(j)}[\frac{\partial h_j}{\partial \theta_r} | S(\theta_1)] | S(\theta_1)\}. \end{aligned}$$

Also

$$\begin{aligned} \langle s, g_r \rangle_{S(\theta_1)} &= \sum_{j=1}^k E[q_{jr} s h_j | S(\theta_1)] \\ &= \sum_{j=1}^k E\{q_{jr} E_{(j)}[s h_j | S(\theta_1)] | S(\theta_1)\} \\ &= \sum_{j=1}^k E\{q_{jr} E_{(j)}[\frac{\partial h_j}{\partial \theta_r} | S(\theta_1)] | S(\theta_1)\}. \end{aligned}$$

Thus g^* is the orthogonal projection of the score function into L_0 .

Once again (b) and (c) follow from part (a) and Theorem 2.2.4.

CHAPTER 4

CONVEXITY AND ITS APPLICATIONS TO STATISTICS

4.1 Introduction

In this chapter, we first prove some general results about convexity, and then apply the results to various statistical problems, which include the theory of optimum experimental designs, the fundamental theorem of mixture distributions due to Lindsay (1983a), and the asymptotic minimaxity of robust estimation due to Huber. Huber (1964) proved an asymptotic minimaxity result for estimating functions about the location parameter. In this chapter, this fundamental result will be generalized to general estimating functions. The geometric optimality of estimating functions proved in Chapter 2 will be used to prove a necessary and sufficient condition for the asymptotic minimaxity of estimating functions in multi-dimensional parameter spaces.

The contents of this chapter are organized as follows: in Section 4.2, a few simple results about matrix valued convex functions will be proved. Also we include some of the well known results in convex analysis, such as the Krein-Milman theorem about extreme points of convex sets, and the Caratheodory theorem about the representation of elements of a convex set in a finite dimensional vector space. In Section 4.3, the results of Section 4.2 are applied to the theory of optimum experimental designs. The fundamental result on optimal design theory is generalized to the matrix valued case. In Section 4.4, the results of Section 4.2 are applied to the mixture distribution

situation; the fundamental result about mixture distribution due to Lindsay (1983a) is an easy consequence. In Section 4.5, the results of Section 4.2 and Chapter 2 will be used to generalize the classical asymptotic minimaxity result of Huber (1964) in the estimating function framework.

4.2 Some Simple Results About Convexity

Let L be a linear space. A subset C of L is said to be convex if for every $x, y \in C$, $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda)y \in C.$$

A function $f : C \rightarrow R$ is said to be convex if for any $x, y \in C$, $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

A symmetric matrix-valued function $N : C \rightarrow M_{k \times k}$ (i. e., for any $x \in C$, $N(x)$ is a symmetric $k \times k$ matrix) is said to be convex, if for any $x, y \in C$, $\lambda \in [0, 1]$,

$$N(\lambda x + (1 - \lambda)y) \leq \lambda N(x) + (1 - \lambda)N(y),$$

where for two $k \times k$ matrices A, B , $A \leq B$ means that $B - A$ is nonnegative (n. n. d.). In the following, we only study properties of matrix valued convex functions, since for $k = 1$, they are reduced to the real valued case.

For every $x, y \in C$, consider the function on $[0, 1]$ as follows

$$N(\lambda; x, y) = N((1 - \lambda)x + \lambda y),$$

then $N(\lambda; x, y)$ is a convex function on λ . The directional derivative of N at x in the direction of y is defined as

$$F_N(x; y) = \lim_{\lambda \rightarrow 0^+} \frac{N(\lambda; x, y) - N(0; x, y)}{\lambda}. \quad (4.2.1)$$

The existence of the limit is justified as follows: Since

$$\lambda_1 = (1 - \frac{\lambda_1}{\lambda_2})0 + \frac{\lambda_1}{\lambda_2}\lambda_2,$$

for $0 < \lambda_1 < \lambda_2 < 1$,

$$N(\lambda_1; x, y) \leq (1 - \frac{\lambda_1}{\lambda_2})N(0; x, y) + \frac{\lambda_1}{\lambda_2}N(\lambda_2; x, y).$$

This implies

$$\frac{N(\lambda_1; x, y) - N(0; x, y)}{\lambda_1} \leq \frac{N(\lambda_2; x, y) - N(0; x, y)}{\lambda_2}, \quad (4.2.2)$$

that is $\frac{N(\lambda; x, y) - N(0; x, y)}{\lambda}$ is a nonincreasing function of λ in $(0, 1]$. Hence the limit in (4.2.1) is well defined.

From (4.2.1) and (4.2.2), for a convex function N ,

$$F_N(x; y) \leq N(1; x, y) - N(0; x, y) = N(y) - N(x), \quad \text{for all } x, y \in C. \quad (4.2.3)$$

The following result will be used repeatedly in the sequel.

Theorem 4.2.1. Suppose that N is convex, then for $x_0 \in C$ satisfies that $N(x_0) \leq N(y)$ for all $y \in C$ if and only if

$$F_N(x_0; y) \geq 0, \quad (4.2.4)$$

for all $y \in C$.

Proof. Suppose that $N(x_0) \leq N(y)$ for all $y \in C$. Then $\frac{N(\lambda; x_0, y) - N(0; x_0, y)}{\lambda}$ is non-negative definite. Hence

$$F_N(x_0; y) = \lim_{\lambda \rightarrow 0^+} \frac{N(\lambda; x_0, y) - N(0; x_0, y)}{\lambda},$$

is n. n. d., for all $y \in C$.

Conversely, if $F_N(x_0; y) \geq 0$ for all $y \in C$, then from (4.2.3),

$$N(y) - N(x_0) \geq F_N(x_0; y) \geq 0.$$

Thus $N(x_0) \leq N(y)$ for all $y \in C$.

Next let L be a locally convex vector space, and let N be a symmetric matrix valued function; then N is said to be Gateaux differentiable at x , if there exists a continuous linear operator $A : L \rightarrow M_{k \times k}$ such that

$$F_N(x; y) = A(y - x), \quad \text{for all } y \in C. \quad (4.2.5)$$

Before stating the next result, let us recall one of the well known results from functional analysis.

Theorem 4.2.2 (Krein-Milman). Let L be a locally convex vector space, and let C be a convex compact subset of L . Then

$$C = \overline{\text{conv}}[\text{ext}(C)],$$

where $\text{ext}(C)$ denotes the set of extreme points of C , and $\overline{\text{conv}}(A)$ denotes the closed convex hull of A , it is the smallest closed convex set containing A .

Now equipped with Gateaux differentiability and Krein-Milman theorem, we are in the position to prove the following result.

Theorem 4.2.3. Let L be a locally convex vector space, and let C be a convex compact subset of L . If N is convex Gateaux differentiable at x_0 , then $x_0 \in C$ satisfies $N(x_0) \leq N(y)$ for all $y \in C$ if and only if

$$F_N(x_0; y) \geq 0, \quad (4.2.6)$$

for all $y \in \text{ext}(C)$.

Proof. Since $F_N(x_0; y) = A(y - x_0)$ for some continuous linear operator A for all $y \in C$, from the definition of Gateaux differentiability, $F_N(x_0; y) \geq 0$, for all $y \in C$ is equivalent to $F_N(x_0; y) \geq 0$, for all $y \in \text{ext}(C)$. Thus Theorem 4.2.3 follows from Theorem 4.2.1.

Next the famous theorem of Caratheodory about the representation of elements of convex set in a finite dimensional vector space is presented. The present proof, taken directly from Silvey (1980), is included for the sake of completeness.

Theorem 4.2.4 (Caratheodory). Let S be a subset of R^n . Then every element c in $\text{conv}(S)$ can be expressed as a convex combination of at most $n + 1$ elements of S . If c is in the boundary of $\text{conv}(S)$, $n + 1$ can be replaced by n .

Proof. Let

$$S' = \{(1, x) : x \in S\}$$

be a subset of R^{n+1} , let K be the convex cone generated by S' . Let $y \in K$, then y can be written as

$$y = \lambda_1 y_1 + \dots + \lambda_m y_m,$$

where each $\lambda_i > 0$ and each $y_i \in S'$. Suppose that the y_i are not linearly independent. Then there exists μ_1, \dots, μ_m , not all zeroes such that

$$\mu_1 y_1 + \dots + \mu_m y_m = 0.$$

Since the first component of each y_i is 1, so $\mu_1 + \dots + \mu_m = 0$. Hence, at least one μ_i is positive. Let λ be the largest number such that $\lambda \mu_i \leq \lambda_i$, $i = 1, \dots, m$; λ is finite since at least one μ_i is positive. Now let $\lambda'_i = \lambda_i - \lambda \mu_i$, then

$$y = \lambda'_1 y_1 + \dots + \lambda'_m y_m,$$

and at least one $\lambda'_i = 0$. Thus y can be expressed as a positive linear combination of fewer than m elements of S' . This argument can continue, until y has been expressed as a positive linear combination of at most $n + 1$ elements of S' , since more than $n + 1$ elements are linearly dependent. Now the first part of the theorem follows by applying the above result to $(1, c) \in S'$.

Next suppose that $y \in K$ and

$$y = \lambda_1 y_1 + \dots + \lambda_{n+1} y_{n+1},$$

where each $\lambda_i > 0$ and the y_i are linearly independent. Then y is an interior point of K . Thus any boundary point of K can be expressed as a positive linear combination

of at most n linearly independent elements of S' . So the second part of the theorem follows.

Proposition 4.2.1. If C is a compact subset of a locally convex vector space, then $\text{conv}(C)$ is compact.

Theorem 4.2.5. (a) With the same notation as Theorem 4.2.3, and the Gateaux differentiability of N on C . The following are equivalent:

- (i) x_0 minimizes $N(x)$;
- (ii) x_0 maximizes $\inf_{y \in C} a^t F_N(x; y)a$, for any $k \times 1$ real vector a ;
- (iii) $\inf_{y \in C} a^t F_N(x_0; y)a = 0$, for any $k \times 1$ real vector a .
- (b) If x_0 minimizes $N(x)$, then (x_0, x_0) is a saddle point of F_N , that is,

$$F_N(x_0; y_1) \geq 0 = F_N(x_0; x_0) \geq F_N(y_2; x_0),$$

for all $y_1, y_2 \in C$.

- (c) If x_0 minimizes $N(x)$, then the support of x_0 is contained in $\{y : F_N(x_0; y) = 0\}$.

More precisely,

$$\{y_i \in C, x_0 = \sum_i \lambda_i y_i, \lambda_i > 0, \sum_i \lambda_i = 1\} \subset \{y : F_N(x_0; y) = 0\}.$$

Proof. (a) First note that from Gateaux differentiability of N , for any real $k \times 1$ vector a , and $x \in C$,

$$\inf_{y \in \text{ext}(C)} a^t F_N(x; y)a = \inf_{y \in C} a^t F_N(x; y)a,$$

and

$$\inf_{y \in C} a^t F_N(x; y)a \leq a^t F_N(x; x)a = 0.$$

((i) \iff (iii)). Note that x_0 minimizes $N(x)$, if and only if for any real $k \times 1$ vector a , and $y \in C$, $a^t F_N(x_0; y)a \geq 0$. The last inequality holds if and only if $\inf_{y \in C} a^t F_N(x_0; y)a \geq 0$, for any $k \times 1$ real vector a . This, in turn, is equivalent to $\inf_{y \in \text{ext}(C)} a^t F_N(x_0; y)a = 0$, for every $k \times 1$ real vector a .

((ii) \iff (iii)). Note that x_0 maximizes $\inf_{y \in C} a^t F_N(x; y)a$, for every $k \times 1$ real vector a , if and only if $\inf_{y \in C} a^t F_N(x_0; y)a \geq 0$, for any $k \times 1$ real vector a . This is equivalent to $\inf_{y \in \text{ext}(C)} a^t F_N(x_0; y)a = 0$, for every $k \times 1$ real vector a .

(b) This follows from Theorem 4.2.1 and the definition of F_N .

(c) If $x_0 = \sum_i \lambda_i y_i$, $\lambda_i > 0$, $\sum_i \lambda_i = 1$, since N is Gateaux differentiable,

$$\begin{aligned} 0 &= F_N(x_0; x_0) = F_N(x_0; \sum_i \lambda_i y_i) \\ &= \sum_i \lambda_i F_N(x_0; y_i). \end{aligned}$$

Since $F_N(x_0; y) \geq 0$ for all $y \in C$, $F_N(x_0; y_i) = 0$ for all y_i .

4.3 Theory of Optimum Experimental Designs

In this section, the results of the previous section are applied to fixed optimal experimental designs. First, we formulate the problem.

Let $f = (f_1, \dots, f_m)$ denote m linearly independent continuous functions on a compact set \mathcal{X} , and let $\theta = (\theta_1, \dots, \theta_m)$ denote a vector of parameters. For each $x \in \mathcal{X}$, an experiment is performed. The outcome is a random variable $y(x)$ with mean value $f(x)^t \theta = \sum_{i=1}^m f_i(x) \theta_i$, and a variance σ^2 , independent of x . The functions f_1, \dots, f_m , called the regression functions, are assumed to be known, while $\theta = (\theta_1, \dots, \theta_m)$ and σ are unknown. An experimental design is a probability measure μ defined on a fixed σ -algebra of subsets of \mathcal{X} , which include the one point subsets. In practice, the experimenter is allowed N uncorrelated observations and the number of observations that he (or she) takes at each $x \in \mathcal{X}$ is proportional to the measure μ . For a given μ , let

$$M(\mu) = ((m_{ij}(\mu)))_{i,j=1}^m, \quad m_{ij}(\mu) = \int_{\mathcal{X}} f_i(x) f_j(x) d\mu(x). \quad (4.3.7)$$

The matrix $M(\mu)$ is called the information matrix of the design μ .

Let \mathcal{H} denote the set of all probability measures on \mathcal{X} with the fixed σ -algebra, and $\mathcal{M} = \{M(\mu) : \mu \in \mathcal{H}\}$, $\phi : \mathcal{M} \longrightarrow M_{k \times k}$ be a symmetric matrix-valued function. The

problem of interest is to determine μ_* , which maximizes $\phi(M(\mu))$ over all probability measures. Any such μ_* will be called ϕ -optimal.

Proposition 4.3.1.

$$\mathcal{M} = \text{conv}(\{f(x)f(x)^t : x \in \mathcal{X}\}).$$

Proof. Since \mathcal{M} is a convex set, and $\{f(x)f(x)^t : x \in \mathcal{X}\} \subset \mathcal{M}$, so

$$\text{conv}(\{f(x)f(x)^t : x \in \mathcal{X}\}) \subset \mathcal{M}.$$

Next since \mathcal{X} is compact, and f is continuous, thus

$$\{f(x)f(x)^t : x \in \mathcal{X}\} \subset \mathcal{M}$$

is compact. Hence

$$\text{conv}(\{f(x)f(x)^t : x \in \mathcal{X}\}) = \overline{\text{conv}}(\{f(x)f(x)^t : x \in \mathcal{X}\}).$$

Also since $\mathcal{M} \subset \overline{\text{conv}}(\{f(x)f(x)^t : x \in \mathcal{X}\})$, hence

$$\mathcal{M} = \text{conv}(\{f(x)f(x)^t : x \in \mathcal{X}\}).$$

From the above proposition and Caratheodory's theorem, the following is true.

Corollary 4.3.1. For any $M(\mu) \in \mathcal{M}$, there exists $x_i \in \mathcal{X}, i = 1, \dots, I, I \leq$

$\frac{m(m+1)}{2} + 1$, such that

$$M(\mu) = \sum_{i=1}^I \lambda_i f(x_i)f(x_i)^t,$$

where $\lambda_i > 0, \sum_{i=1}^I \lambda_i = 1$. If $M(\mu)$ is a boundary point of \mathcal{M} , the inequality involving I can be reduced to $I \leq \frac{m(m+1)}{2}$.

From the practical point of view, this corollary is extremely important. For it means that if ϕ is maximal at M_* , then M_* can always be expressed as $M(\mu_*)$, where μ_* is a discrete design measure supported by at most $\frac{m(m+1)}{2} + 1$ points.

Now we are in the position to prove the fundamental theorem in optimum design theory, which is a generalization of the result in Silvey (1980) to the matrix valued case.

Theorem 4.3.1. (A) If ϕ is a concave function on \mathcal{M} , then $M(\mu_*)$ is ϕ -optimal if and only if

$$F_\phi(M(\mu_*), M(\mu)) \leq 0,$$

for all $\mu \in H$;

(B) If ϕ is a concave function on \mathcal{M} , which is Gateaux differentiable on \mathcal{M} , then $M(\mu_*)$ is ϕ -optimal if and only if

$$F_\phi(M(\mu_*), f(x)f(x)^t) \leq 0,$$

for all $x \in \mathcal{X}$;

(C) If ϕ is Gateaux differentiable at $M(\mu_*)$, and $M(\mu_*)$ is ϕ optimal, then

$$\begin{aligned} & \{x_i \in \mathcal{X} : M(\mu_*) = \sum_i \lambda_i f(x_i)f(x_i)^t, \lambda_i > 0, \sum_i \lambda_i = 1\} \\ & \subset \{x \in \mathcal{X} : F_\phi(M(\mu_*), f(x)f(x)^t) = 0\}. \end{aligned}$$

Proof. They are easy consequences of Theorem 4.2.1 - 4.2.3.

Next we apply Theorem 4.3.1 to study the relationship between D and G optimal designs.

The D -optimality criterion is defined by the criterion function

$$\begin{aligned} \phi[M(\mu)] &= \log \det[M(\mu)], & \text{if } \det[M(\mu)] \neq 0 \\ &= -\infty, & \text{if } \det[M(\mu)] = 0. \end{aligned} \quad (4.3.8)$$

$\mu_* \in \mathcal{H}$ is said to be D -optimal if μ_* maximizes ϕ . Let \mathcal{M} denote the set of all positive definite matrices, then ϕ has the following properties:

- (a) ϕ is continuous on \mathcal{M} ;
- (b) ϕ is concave on \mathcal{M} ;

(c) ϕ is Gateaux differentiable at M_1 if it is nonsingular, and

$$F_\phi(M_1, M_2) = \text{tr}(M_2 M_1^{-1}) - k.$$

Proof. (a) The continuity of ϕ follows from the continuity of \det .

(b) We want to show that, for every $\lambda \in (0, 1)$, and $M_1, M_2 \in \mathcal{M}$,

$$\phi[(1 - \lambda)M_1 + \lambda M_2] \geq (1 - \lambda)\phi(M_1) + \lambda\phi(M_2).$$

This inequality is obvious if either M_1 or M_2 is singular. Thus we only need to prove the inequality if both M_1 and M_2 are nonsingular. From a standard result from matrix algebra, there is a nonsingular matrix U such that

$$U M_1 U^t = I, \quad U M_2 U^t = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_k).$$

Using the concavity of \log ,

$$\begin{aligned} \phi[(1 - \lambda)M_1 + \lambda M_2] &= \log \det\{U^{-1}[(1 - \lambda)I + \lambda\Lambda]U^{-1t}\} \\ &\geq \log \det U^{-2} + \sum_{i=1}^k \lambda \log \lambda_i \\ &= (1 - \lambda) \log \det U^{-2} + \lambda \log \det(U^{-1}\Lambda U^{-1t}) \\ &= (1 - \lambda)\phi(M_1) + \lambda\phi(M_2). \end{aligned}$$

(c) For nonsingular matrix M_1 , we have

$$\begin{aligned} &\phi(M_1 + \epsilon M_2) - \phi(M_1) \\ &= \log \det(I + \epsilon M_2 M_1^{-1}) \\ &= \log\{1 + \epsilon \text{tr}(M_2 M_1^{-1})\} + O(\epsilon^2) \\ &= \epsilon \text{tr}(M_2 M_1^{-1}) + O(\epsilon^2). \end{aligned}$$

Thus

$$F_\phi(M_1, M_2) = \text{tr}[(M_2 - M_1)M_1^{-1}] = \text{tr}(M_2 M_1^{-1}) - k. \quad (4.3.9)$$

The G -optimality criterion is defined by the criterion function

$$\begin{aligned}\phi[M(\mu)] &= \max_{x \in \mathcal{X}} f^t(x) M^{-1}(\mu) f(x), & \text{if } \det M(\mu) \neq 0, \\ &= \infty, & \text{if } \det M(\mu) = 0.\end{aligned}\quad (4.3.10)$$

A design μ_* is said to be G -optimal if

$$\phi[M(\mu_*)] \leq \phi[M(\mu)], \quad \text{for all } \mu \in \mathcal{H}.$$

Example. The equivalence of D and G optimal designs.

In this example, we derive the famous equivalence theorem due to Kiefer and Wolfowitz (1960) about the D and G optimal designs.

Theorem 4.3.2 (Kiefer and Wolfowitz). If $\mu_* \in \mathcal{H}$ satisfies the condition that $M(\mu_*)$ is nonsingular, then μ_* is D -optimal if and only if μ_* is G -optimal.

Proof. (\Rightarrow). From Theorem 4.3.1 and (4.3.9), μ_* is D -optimal if and only if

$$\text{tr}[f(x)f^t(x)M(\mu_*)^{-1}] \leq k, \quad \text{for all } x \in \mathcal{X},$$

that is

$$\max_{x \in \mathcal{X}} f^t(x) M^{-1}(\mu_*) f(x) \leq k.$$

On the other hand, for any $\mu \in \mathcal{H}$ such that $M(\mu)$ is nonsingular,

$$\begin{aligned}\max_{x \in \mathcal{X}} f^t(x) M^{-1}(\mu) f(x) &\geq \int_{\mathcal{X}} f^t(x) M^{-1}(\mu) f(x) d\mu(x) \\ &= \text{tr}[M^{-1}(\mu) \int_{\mathcal{X}} f(x) f^t(x) d\mu(x)] \\ &= \text{tr}[M^{-1}(\mu) M(\mu)] = k.\end{aligned}$$

Hence,

$$k = \max_{x \in \mathcal{X}} f^t(x) M^{-1}(\mu_*) f(x) \leq \max_{x \in \mathcal{X}} f^t(x) M^{-1}(\mu) f(x),$$

for any $\mu \in \mathcal{H}$ such that $M(\mu)$ is nonsingular, therefore, μ_* is G -optimal.

(\Leftarrow). Now suppose that μ_1 is G -optimal, then from the definition, $M(\mu_1)$ is nonsingular. Let μ_* be any D -optimal design. Then

$$1 \leq \det[M(\mu_*)M^{-1}(\mu_1)] = \prod_{i=1}^k \lambda_i,$$

where $\lambda_1, \dots, \lambda_k$ are the eigenvalues of the matrix $M^{-1/2}(\mu_1)M(\mu_*)M^{-1/2}(\mu_1)$. Hence

$$\begin{aligned} \left(\prod_{i=1}^k \lambda_i\right)^{1/k} &\leq \frac{1}{k} \sum_{i=1}^k \lambda_i = \frac{1}{k} \operatorname{tr}[M(\mu_*)M^{-1}(\mu_1)] \\ &= \frac{1}{k} \int_{\mathcal{X}} f^t(x) M^{-1}(\mu_1) f(x) d\mu_*(x) \\ &\leq \frac{1}{k} \max_{x \in \mathcal{X}} f^t(x) M^{-1}(\mu_1) f(x) = 1. \end{aligned}$$

Therefore,

$$\det M(\mu_1) = \det M(\mu_*),$$

hence μ_1 is D -optimal.

4.4 Fundamental Theorem of Mixture Distributions

In this section, we apply the results of Section 2 to the mixture distribution problem. The fundamental result is due to Lindsay (1982, 1995). We begin with the formulation of the problem.

Let $\{f_\theta : \theta \in \Theta\}$ be a parametric family of densities with respect to some σ -finite measure, let the parameter space Θ have a σ -algebra of measurable sets which contains all atomic sets $\{\theta\}$. Let \mathcal{H} be the class of all probability measures on Θ . Define the function

$$f_Q(x) = \int_{\Theta} f_\theta(x) dQ(\theta), \quad Q \in \mathcal{H}, \quad (4.4.11)$$

to be the mixture density corresponding to mixing distribution Q . Since the densities $\{f_\theta\}$ correspond to the atomic mixing distribution $\{\delta(\theta)\}$, which assign probability one to any set containing θ , they are called the atomic densities. A finite discrete mixing distribution with support size J will be expressed as $Q = \sum_j \pi_j \delta(\theta_j)$, and the θ_j 's are distinct, $\pi_j > 0$, $\sum_j \pi_j = 1$.

Given a random sample X_1, \dots, X_n from the mixture density f_Q , the objective will be to estimate the mixing distribution Q by \hat{Q}_n , a maximizer of the likelihood

$$L(Q) = \prod_{i=1}^n f_Q(x_i).$$

Now suppose that the observation vector (x_1, \dots, x_n) has K distinct data points y_1, \dots, y_K , and let n_k be the number of x 's which equals to y_k . Define the atomic and mixture likelihood to be $f_\theta = (f_\theta(y_1), \dots, f_\theta(y_K))$, and $f_Q = (f_Q(y_1), \dots, f_Q(y_K))$, respectively. The likelihood curve is the function from Θ to R defined by $\theta \rightarrow f_\theta$. The orbit of this curve, given by $\Gamma = \{f_\theta : \theta \in \Theta\}$, represents all possible fitted values of the atomic likelihood vector. Then $\text{conv}(\Gamma) = \{f_Q : Q \in \mathcal{H}, |\text{support}(Q)| < \infty\}$, where $|A|$ denotes the cardinality of A . Furthermore, if Θ is compact and f_θ is a continuous function of θ , then $\text{conv}(\Gamma) = \{f_Q : Q \in \mathcal{H}\}$. In this case, maximizing $L(Q)$ over $Q \in \mathcal{H}$ may be accomplished by maximizing the concave functional $\phi(f) = \sum_k n_k \log f_k$ over f in the K -dimensional set $\text{conv}(\Gamma)$. Note that $\phi(f)$ is a strict concave function of f .

Now we are in the position to state the fundamental result about mixture distributions.

Theorem 4.4.1 (Lindsay). Suppose that Θ is compact, and f_θ is continuous.

(A) There exists a unique vector \hat{f} on the boundary of $\text{conv}(\Gamma)$ which maximizes the log likelihood $\phi(f)$ on $\text{conv}(\Gamma)$. \hat{f} can be expressed as f_Q , where Q has K or fewer points of support.

(B) The measure \hat{Q} which maximizes $\log L(Q)$ can be equivalently characterized by three conditions:

- (i) \hat{Q} maximizes $L(Q)$;
- (ii) \hat{Q} minimizes $\sup_{\theta} D(\theta; Q)$;
- (iii) $\sup_{\theta} D(\theta; \hat{Q}) = 0$.

(C) The point (\hat{f}, \hat{f}) is a saddle point of Φ , in the sense that

$$\Phi(f_{Q_0}, \hat{f}) \leq 0 = \Phi(\hat{f}, \hat{f}) \leq \Phi(\hat{f}, f_{Q_1}),$$

for all $Q_0, Q_1 \in H$.

(D). The support of \hat{Q} is contained in the set of θ for which $D(\theta, \hat{Q}) = 0$.

Proof. The results are easy consequences of Caratheodory's theorem and Theorem 4.2.5.

4.5 Asymptotic Minimality of Estimating Functions

In this section, the famous asymptotic minimality result due to Huber (1964) will be generalized. First we formulate the problem of interest.

Let Θ be an open subset of R^k , \mathcal{X} be the sample space, a function

$$g : \mathcal{X} \times \Theta \longrightarrow R^k,$$

is called an unbiased estimating function if

$$E[g(X; \theta) | \theta] = 0,$$

for all $\theta \in \Theta$. An unbiased estimating function is called regular if

$$E\left[\frac{\partial g}{\partial \theta} | \theta\right],$$

is nonsingular for all $\theta \in \Theta$.

In the rest of this section, the regularity conditions (I) - (IV) of Section 3.3.1 for estimating functions will always be assumed.

Let C be a convex set of distribution functions such that every $F \in C$ has an absolutely continuous density f satisfying

$$I(F) = E\left[\left(\frac{\partial \log f}{\partial \theta}\right)\left(\frac{\partial \log f}{\partial \theta}\right)^t | F, \theta\right], \quad (4.5.12)$$

is positive definite. Let L be the space of unbiased estimating functions with respect to C , that is every element of L is unbiased with respect to every distribution in C . Let Φ_0 be the subset of L which consists of all regular unbiased estimating functions in L .

Consider the function $K : \Phi_0 \times C \longrightarrow M_{k \times k}$, defined by

$$K(\phi, F) = E\left[\frac{\partial \phi}{\partial \theta} | F, \theta\right]^t (E[\phi \phi^t | F, \theta])^{-1} E\left[\frac{\partial \phi}{\partial \theta} | F, \theta\right], \quad (4.5.13)$$

for all $\phi \in \Phi_0, F \in C$. Note that when $k = 1$, then

$$K(\phi, F) = \frac{(\int_X \phi f' dx)^2}{\int_X \phi^2 f dx},$$

For every $F \in C$, for any $g_1, g_2 \in L$, the inner product of g_1 and g_2 is defined by

$$\langle g_1, g_2 \rangle_F = E[g_1 g_2^t | F].$$

For every $F \in C$, the orthogonal projection of the score function of F into the subspace L_0 , with respect to the inner product $\langle \cdot, \cdot \rangle_F$ (if it exists), is denoted by ϕ_F .

Lemma 4.5.1. (a) For any $(u, v) \in R \times R^+$, the function defined by

$$h(u, v) = \frac{u^2}{v},$$

is convex, that is for any $(u_i, v_i) \in R \times R^+, i = 1, 2, \lambda \in (0, 1)$

$$\frac{(\lambda u_1 + (1 - \lambda)u_2)^2}{\lambda v_1 + (1 - \lambda)v_2} \leq \lambda \frac{u_1^2}{v_1} + (1 - \lambda) \frac{u_2^2}{v_2};$$

(b) For any $(M_1, M_2) \in M_{k \times k} \times M_{k \times k}^+$, where $M_{k \times k}^+$ denotes the set of all $k \times k$ positive definite matrices, the matrix valued function defined by

$$J(M_1, M_2) = M_1^t M_2^{-1} M_1,$$

is convex in the sense that, for any $(M_1, M_2), (M_3, M_4) \in M_{k \times k} \times M_{k \times k}^+, \lambda \in (0, 1)$,

$$J(\lambda) = [\lambda M_1 + (1 - \lambda)M_3]^t [\lambda M_2 + (1 - \lambda)M_4]^{-1} [\lambda M_1 + (1 - \lambda)M_3],$$

is convex in λ .

Proof. (a)

$$\frac{\partial h}{\partial u} = \frac{2u}{v}, \quad \frac{\partial h}{\partial v} = -\frac{u^2}{v^2},$$

$$\frac{\partial^2 h}{\partial^2 u} = \frac{2}{v}, \quad \frac{\partial^2 h}{\partial u \partial v} = -\frac{2u}{v^2}, \quad \frac{\partial^2 h}{\partial^2 v} = \frac{2u^2}{v^3}.$$

The matrix

$$\begin{bmatrix} 2/v & -2u/v^2 \\ -2u/v^2 & 2u^2/v^3 \end{bmatrix}.$$

is non negative definite, so h is convex.

(b) By straightforward calculation, and using repeatedly the relation

$$\frac{dM^{-1}}{d\lambda} = -M^{-1} \frac{dM}{d\lambda} M^{-1},$$

one gets,

$$\frac{dJ(\lambda)}{d\lambda} = (M_1 - M_3)^t [\lambda M_2 + (1 - \lambda)M_4]^{-1} [\lambda M_1 + (1 - \lambda)M_3]$$

$$- [\lambda M_1 + (1 - \lambda)M_3]^t [\lambda M_2 + (1 - \lambda)M_4]^{-1} (M_2 - M_4) [\lambda M_2 + (1 - \lambda)M_4]^{-1} [\lambda M_1 + (1 - \lambda)M_3]$$

$$+[\lambda M_1 + (1 - \lambda)M_3]^t[\lambda M_2 + (1 - \lambda)M_4]^{-1}(M_1 - M_3), \quad (4.5.14)$$

and

$$\begin{aligned} \frac{d^2 J(\lambda)}{d^2 \lambda} &= 2\{(M_1 - M_3)^t[\lambda M_2 + (1 - \lambda)M_4]^{-1}(M_1 - M_3) \\ &\quad [\lambda M_1 + (1 - \lambda)M_3]^t[\lambda M_2 + (1 - \lambda)M_4]^{-1}(M_2 - M_4)[\lambda M_2 + (1 - \lambda)M_4]^{-1} \\ &\quad (M_2 - M_4)[\lambda M_2 + (1 - \lambda)M_4]^{-1}[\lambda M_1 + (1 - \lambda)M_3] \\ &\quad -(M_1 - M_3)^t[\lambda M_2 + (1 - \lambda)M_4]^{-1}(M_2 - M_4)[\lambda M_2 + (1 - \lambda)M_4]^{-1}[\lambda M_1 + (1 - \lambda)M_3] \\ &\quad [\lambda M_1 + (1 - \lambda)M_3]^t[\lambda M_2 + (1 - \lambda)M_4]^{-1}(M_2 - M_4)[\lambda M_2 + (1 - \lambda)M_4]^{-1}(M_1 - M_3)\} \\ &= 2(AA^t + B^t B - AB - B^t A^t) \\ &= 2(A - B^t)(A - B^t)^t \geq 0, \end{aligned} \quad (4.5.15)$$

where

$$A = (M_1 - M_3)^t[\lambda M_2 + (1 - \lambda)M_4]^{-1/2},$$

$$B = [\lambda M_2 + (1 - \lambda)M_4]^{-1/2}(M_2 - M_4)[\lambda M_2 + (1 - \lambda)M_4]^{-1}[\lambda M_1 + (1 - \lambda)M_3].$$

This completes the proof of the Lemma.

Note that part (a) of Lemma 4.5.1 was proved by Huber (1964) by using a different argument. Also from part (b),

$$\frac{dJ(0)}{d\lambda} = (M_1 - M_3)^t M_4^{-1} M_3 - M_3^t M_4^{-1} (M_2 - M_4) M_4^{-1} M_3 + M_3^t M_4^{-1} (M_1 - M_3). \quad (4.5.16)$$

We will use this identity in Section 6.2.

4.5.1 One Dimensional Case

In this subsection, a necessary and sufficient condition of the asymptotic minimaxity of estimating functions will be given when the parameter space is one-dimensional. This result generalizes Theorem 2 of Huber (1964).

Theorem 4.5.1. Suppose the parameter space is one dimensional. Then (ϕ_{F_0}, F_0) is a saddle point of K , that is

$$K(\phi, F_0) \leq K(\phi_{F_0}, F_0) \leq K(\phi_{F_0}, F),$$

for all $\phi \in \Phi$, and $F \in C$, if and only if

$$\int_{\mathcal{X}} (2\phi_{F_0}(f' - f'_0) - (\phi_{F_0})^2(f - f_0))dx \geq 0, \quad (4.5.17)$$

where f' denotes the derivative of f with respect to the parameter.

Proof. Note that since ϕ_{F_0} is the orthogonal projection of s_{F_0} into L_0 ,

$$K(\phi, F_0) \leq K(\phi_{F_0}, F_0),$$

for all $\phi \in \Phi$. This fact has been established in Chapter 2.

Also for any $F_1 \in C$, consider the function

$$h_{F_1} : [0, 1] \longrightarrow R,$$

given by

$$h_{F_1}(t) = \frac{(\int_{\mathcal{X}} \phi_{F_0}[(1-t)f'_0 + tf'_1]dx)^2}{\int_{\mathcal{X}} \phi_{F_0}^2[(1-t)f_0 + tf_1]dx}. \quad (4.5.18)$$

Then by (a) of Lemma 4.5.1, h_{F_1} is a convex function, and by direct calculation,

$$\begin{aligned} h'_{F_1}(0+) &= \frac{\int_{\mathcal{X}} \phi_{F_0} f'_0 dx}{(\int_{\mathcal{X}} \phi_{F_0}^2 f_0 dx)^2} [2 \int_{\mathcal{X}} \phi_{F_0} g' dx \int_{\mathcal{X}} \phi_{F_0}^2 f_0 dx \\ &\quad - \int_{\mathcal{X}} \phi_{F_0} f'_0 dx \int_{\mathcal{X}} \phi_{F_0}^2 g dx], \end{aligned} \quad (4.5.19)$$

where $g = f_1 - f_0$. Since ϕ_{F_0} is the orthogonal projection of s_{F_0} into L_0 with respect to the inner product $\langle \cdot, \cdot \rangle_{F_0}$,

$$\int_{\mathcal{X}} \phi_{F_0} f'_0 dx = \int_{\mathcal{X}} \phi_{F_0} \frac{f'_0}{f_0} f_0 dx = \int_{\mathcal{X}} \phi_{F_0}^2 f_0 dx.$$

Hence,

$$h'_{F_1}(0+) = \int_{\mathcal{X}} (2\phi_{F_0}g' - \phi_{F_0}^2g)dx. \quad (4.5.20)$$

Only if. Suppose that (ϕ_{F_0}, F_0) is a saddle point of K . Then for any $F_1 \in C$, and every $t \in (0, 1)$,

$$h_{F_1}(0) = K(\phi_{F_0}, F_0) \leq K(\phi_{F_0}, (1-t)F_0 + tF_1) = h_{F_1}(t).$$

Now, since $h'_{F_1}(0+) \geq 0$,

$$\int_{\mathcal{X}} (2\phi_{F_0}g' - \phi_{F_0}^2g)dx \geq 0,$$

where $g = f_1 - f_0$.

If. Suppose that

$$\int_{\mathcal{X}} (2\phi_{F_0}g' - \phi_{F_0}^2g)dx \geq 0,$$

where $g = f_1 - f_0$. Then from Theorem 4.2.1, h_{F_1} is a monotone function in $[0, 1]$.

Hence,

$$h_{F_1}(0) = K(\phi_{F_0}, F_0) \leq h_{F_1}(1) = K(\phi_{F_0}, F_1).$$

Thus (ϕ_{F_0}, F_0) is a saddle point of K . This completes the proof of Theorem 4.5.1.

Corollary 4.5.1 (Huber) . Assume that $F_0 \in C$ such that $I(F_0) \leq I(F)$ for all $F \in C$, and $\phi_0 = \frac{f'_0}{f_0} \in \Phi$. Then (ϕ_0, F_0) is a saddle point of K .

Proof. For any $F_1 \in C$, consider the function

$$h_{F_1}(t) = I((1-t)F_0 + tF_1) = \int_{\mathcal{X}} \frac{(f'_0 + t(f_0 - f_1)')^2}{f_0 + t(f_1 - f_0)} dx.$$

Then by (a) of Lemma 4.5.1, h_{F_1} is convex, and attains its minimum at $t = 0$. Thus

$$0 \leq h'_{F_1}(0+) = \int_{\mathcal{X}} [2\frac{f'_0}{f_0}g' - (\frac{f'_0}{f_0})^2g]dx, \quad (4.5.21)$$

where $g = f_1 - f_0$. The above equality follows from the Lebesgue dominated convergence theorem, and the facts that

$$\frac{1}{t} \left[\frac{(f'_t)^2}{f_t} - \frac{(f'_0)^2}{f_0} \right] \rightarrow 2 \frac{f'_0}{f_0} g' - \left(\frac{f'_0}{f_0} \right)^2 g,$$

and

$$\frac{1}{t} \left[\frac{(f'_t)^2}{f_t} - \frac{(f'_0)^2}{f_0} \right] \leq \frac{(f'_1)^2}{f_1} - \frac{(f'_0)^2}{f_0},$$

uniformly in $t \in (0, 1)$.

4.5.2 Multi-Dimensional Case

In this subsection, by using the geometry of optimal estimating functions proved in Chapter 2, a necessary and sufficient condition of the asymptotic minimaxity result for estimating functions in a multi dimensional parameter space will be given. This result generalizes one main result of Huber (1964) to the multi dimensional parameter space.

Theorem 4.5.2. Suppose the parameter space is multi-dimensional. Then (ϕ_{F_0}, F_0) is a saddle point of K , that is

$$K(\phi, F_0) \leq K(\phi_{F_0}, F_0) \leq K(\phi_{F_0}, F),$$

for all $\phi \in \Phi$, and $F \in C$, if and only if

$$\int_{\mathcal{X}} [\phi_{F_0} \left(\frac{\partial f}{\partial \theta} - \frac{\partial f_0}{\partial \theta} \right)^t + \left(\frac{\partial f}{\partial \theta} - \frac{\partial f_0}{\partial \theta} \right) (\phi_{F_0})^t - \phi_{F_0} \phi_{F_0}^t (f - f_0)] dx \quad (4.5.22)$$

is non negative definite.

Proof. Note that since ϕ_{F_0} is the orthogonal projection of s_{F_0} into L_0 ,

$$K(\phi, F_0) \leq K(\phi_{F_0}, F_0),$$

for all $\phi \in \Phi$. This has been proved in Chapter 2.

Also for any $F_1 \in C$, consider the function

$$J_{F_1} : [0, 1] \longrightarrow M_{k \times k},$$

given by

$$\begin{aligned} J_{F_1}(\lambda) = & \left(\int_{\mathcal{X}} \phi_{F_0} [(1 - \lambda) \frac{\partial f_0}{\partial \theta} + \lambda \frac{\partial f_1}{\partial \theta}]^t dx \right)^t \left(\int_{\mathcal{X}} \phi_{F_0} \phi_{F_0}^t [(1 - \lambda) f_0 + \lambda f_1] dx \right)^{-1} \\ & \left(\int_{\mathcal{X}} \phi_{F_0} [(1 - \lambda) \frac{\partial f_0}{\partial \theta} + \lambda \frac{\partial f_1}{\partial \theta}]^t dx \right). \end{aligned} \quad (4.5.23)$$

From (b) of Lemma 4.5.1, J_{F_1} is convex, and by direct calculation,

$$J_{F_1}(0+) = (M_1 - M_3)^t M_4^{-1} M_3 -$$

$$M_3^t M_4^{-1} (M_2 - M_4) M_4^{-1} M_3 + M_3^t M_4^{-1} (M_1 - M_3),$$

where

$$M_1 = \int_{\mathcal{X}} \phi_{F_0} \left(\frac{\partial f_1}{\partial \theta} \right)^t dx, \quad M_3 = \int_{\mathcal{X}} \phi_{F_0} \left(\frac{\partial f_0}{\partial \theta} \right)^t dx,$$

$$M_2 = \int_{\mathcal{X}} \phi_{F_0} \phi_{F_0}^t f_1 dx, \quad M_4 = \int_{\mathcal{X}} \phi_{F_0} \phi_{F_0}^t f_0 dx.$$

Since ϕ_{F_0} is the orthogonal projection of s_{F_0} into L_0 with respect to $\langle \cdot, \cdot \rangle_{F_0}$,

$M_3 = M_4$. Hence,

$$\begin{aligned} J_{F_1}(0+) &= (M_1 - M_3)^t + (M_1 - M_3) - (M_2 - M_4) \\ &= \int_{\mathcal{X}} \left[\phi_{F_0} \left(\frac{\partial f}{\partial \theta} - \frac{\partial f_0}{\partial \theta} \right)^t + \left(\frac{\partial f}{\partial \theta} - \frac{\partial f_0}{\partial \theta} \right) (\phi_{F_0})^t - \phi_{F_0} \phi_{F_0}^t (f - f_0) \right] dx. \end{aligned} \quad (4.5.24)$$

Only if. Suppose that (ϕ_{F_0}, F_0) is a saddle point of K . Then for any $F_1 \in C$, and every $t \in (0, 1)$,

$$J_{F_1}(0) = K(\phi_{F_0}, F_0) \leq K(\phi_{F_0}, (1 - t)F_0 + tF_1) = J_{F_1}(t).$$

Thus from the definition of $J'_{F_1}(0+)$, $J'_{F_1}(0+)$ is non negative definite. Hence,

$$\int_{\mathcal{X}} [\phi_{F_0}(\frac{\partial g}{\partial \theta})^t + \frac{\partial g}{\partial \theta}(\phi_{F_0})^t - \phi_{F_0} \phi_{F_0}^t g] dx,$$

is non negative definite, where $g = f_1 - f_0$.

If. Now suppose that

$$\int_{\mathcal{X}} [\phi_{F_0}(\frac{\partial g}{\partial \theta})^t + \frac{\partial g}{\partial \theta}(\phi_{F_0})^t - \phi_{F_0} \phi_{F_0}^t g] dx$$

is non negative definite, where $g = f_1 - f_0$. Then from Theorem 4.2.1, J_{F_1} is a monotone function in $[0, 1]$. Hence,

$$J_{F_1}(0) = K(\phi_{F_0}, F_0) \leq J_{F_1}(1) = K(\phi_{F_0}, F_1).$$

Hence, (ϕ_{F_0}, F_0) is a saddle point of K . This completes the proof.

Corollary 4.5.2. Assume that $F_0 \in C$ is such that $I(F_0) \leq I(F)$ for all $F \in C$, and $s_{F_0} = \frac{f'_0}{f_0} \in \Phi$. Then (s_{F_0}, F_0) is a saddle point of K .

Proof. For any $F_1 \in C$, consider the function

$$\begin{aligned} J_{F_1}(\lambda) &= I((1 - \lambda)F_0 + \lambda F_1) \\ &= \int_{\mathcal{X}} \frac{\partial(f_0 + \lambda(f_0 - f_1))}{\partial \theta} \left(\frac{\partial(f_0 + \lambda(f_0 - f_1))}{\partial \theta} \right)^t \frac{1}{f_0 + \lambda(f_1 - f_0)} dx. \end{aligned}$$

Then by (b) of Lemma 4.5.1, J_{F_1} is convex, and attains its minimum at $t = 0$. Thus

$$J'_{F_1}(0+) = \int_{\mathcal{X}} [\phi_{F_0}(\frac{\partial g}{\partial \theta})^t + \frac{\partial g}{\partial \theta}(\phi_{F_0})^t - \phi_{F_0} \phi_{F_0}^t g] dx, \quad (4.5.25)$$

is non negative definite, where $g = f_1 - f_0$. The above equality follows from the Lebesgue dominated convergence theorem and the facts that

$$\frac{1}{\lambda} \left[\frac{\partial f_{\lambda}}{\partial \theta} \left(\frac{\partial f_{\lambda}}{\partial \theta} \right)^t - \frac{\partial f_0}{\partial \theta} \left(\frac{\partial f_0}{\partial \theta} \right)^t \right] \rightarrow \frac{\partial f_0}{\partial \theta} \frac{1}{f_0} \left(\frac{\partial g}{\partial \theta} \right)^t + \frac{\partial g}{\partial \theta} \left(\frac{\partial f_0}{\partial \theta} \right)^t \frac{1}{f_0} - \frac{\partial f_0}{\partial \theta} \left(\frac{\partial f_0}{\partial \theta} \right)^t \frac{1}{(f_0)^2} g,$$

and

$$\frac{1}{\lambda} \left[\frac{\frac{\partial f_{\lambda}}{\partial \theta} \left(\frac{\partial f_{\lambda}}{\partial \theta} \right)^t}{f_{\lambda}} - \frac{\frac{\partial f_0}{\partial \theta} \left(\frac{\partial f_0}{\partial \theta} \right)^t}{f_0} \right] \leq \frac{\frac{\partial f_1}{\partial \theta} \left(\frac{\partial f_1}{\partial \theta} \right)^t}{f_1} - \frac{\frac{\partial f_0}{\partial \theta} \left(\frac{\partial f_0}{\partial \theta} \right)^t}{f_0}.$$

CHAPTER 5

SUMMARY AND FUTURE RESEARCH

5.1 Summary

In this dissertation, we have studied optimal estimating functions through the introduction of the generalized inner product space. It turns out that, the orthogonal projection of the score function into the subspace of estimating functions (if it exists), is optimal in that subspace. Also, the estimating function theory in the Bayesian framework is studied. We have shown that the orthogonal projection of the posterior score function into a subspace of estimating functions (if it exists) is optimal in that subspace. The geometry of estimating functions in the presence of nuisance parameters is also studied. The geometric idea of conditional, marginal and partial likelihood inference become transparent when viewed as orthogonal projections of score functions into appropriate subspaces. Finally, a general result about matrix valued convex functions was also proved, and then this result was applied to study optimum experimental designs, mixture distributions and asymptotic minimaxity of estimating functions.

5.2 Future Research

We have studied the geometry of estimating functions in the discrete setting; it will be of great interest to extend these results to the martingale framework. I believe that there is a lot of potential in pursuing a vigorous research in this direction.

In the last decade, there are major advances in the study of geometry of optimum experimental designs. I believe most of these geometric results are direct consequences of the duality theory in convex analysis. As far as I know, the applications of duality theory to statistics are very limited. It will be of great interest to establish a general duality theory in the statistical framework.

BIBLIOGRAPHY

- Amari, S. I. and Kumon, M. (1988), Estimation in the presence of infinitely many nuisance parameters - geometry of estimating functions. *Ann. Statist.*, 16 (3), 1044-1068.
- Bhaskar, V. P. (1972), On a measure of efficiency of an estimating equation. *Sankhya A*, 34, 467-472.
- Bhaskar, V. P. (1989), Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *J. Stat. Plan. Inf.*, 21, 139-160.
- Bhaskar, V. P. (1991a), Loss of information in the presence of nuisance parameters and partial sufficiency. *J. Stat. Plan. Inf.*, 28, 195-203.
- Bhaskar, V. P. (1991b), Sufficiency, ancillarity, and information in estimating functions. Estimating Functions, edited by V. P. Godambe, Oxford University Press, New York, 241-254.
- Bhaskar, V. P. and Srinivasan, C. (1994), On Fisher information inequalities in the presence of nuisance parameters. *Ann. Inst. Stat. Math.*, 46, 593-604.
- Breslow, N. E. and Clayton, D. G. (1993), Approximate inference in generalized linear mixed models. *Journal of American Statistical Association*, 88 (421), 9-25.
- Chaloner, K. and Larntz, K. (1989), Optimal Bayesian design applied to logistic regression experiments. *J. Stat. Plan. Inf.*, 21, 1991-208.
- Cox, D. R. (1972), Regression models and life tables (with discussion). *J. R. Stat. Soc. B*, 34, 187-220.
- Cox, D. R. (1975), Partial likelihood. *Biometrika*, 62, 269-276.
- Crowder, M. (1995), On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, 82, 407-410.
- DasGupta, A. and Studden, W. (1991), Robust Bayesian designs. *Ann. Stat.*,
- Desmond, A. F. (1991), Quasi-likelihood, stochastic processes, and optimal estimating functions. Estimating Functions, edited by V. P. Godambe. Oxford University Press, New

York, 133-146.

- Dette, H. (1993), Elfving's theorem for D -optimality. *Ann. Stat.*, 21 (2), 753-766.
- Dette, H. and Studden, W. J. (1993), Geometry of E -optimality. *Ann. Stat.*, 21 (1), 416-433.
- Diggle, P., Liang, K. Y. and Zeger, S. L. (1994), Analysis of Longitudinal Data. Oxford University Press, New York.
- Durbin, J. (1960). Estimation of parameters in time series regression models. *J. R. Stat. Soc. B*, 22, 139-153.
- Efron, B. and Stein, C (1981) , The jackknife estimate of variance. *Ann. Statist.*, 9 (2), 586-596.
- Elfving, G. (1952), Optimum allocation in linear regression. *Ann. Math. Stat.*, 23, 255-262.
- Elfving, G. (1959), Design of linear experiments. *Cramer Restschrift Volume*. Wiley, New York. 58-58.
- El-Krunz, S. M. and Studden, W. J. (1991), Bayesian optimal designs for linear regression models, *Ann. Statist.*, 19 (4), 2183-2208.
- Ferreira, P. E. (1981), Extending Fisher's measure of information. *Biometrika*, 68, 695-698.
- Ferreira, P. E. (1982), Estimating equations in the presence of prior knowledge. *Biometrika*, 69, 667-669.
- Firth, D. (1987), On the efficiency of quasi-likelihood estimation. *Biometrika*, 74, 233-245.
- Ghosh, M. (1990), On a Bayesian analog of the theory of estimating function. D. G. Khatri Memorial Volume, *Gujarat Stat. Rev.*, 47-52.
- Ghosh, M. and Rao, J. N. K. (1994), Small area estimation: an appraisal (with discussion). *Statistical Sciences*, 9 (1), 55-93.
- Godambe, V. P. (1960). An optimum property of a regular maximum likelihood estimation. *Ann. Math. Stat.*, 31, 1208-1212.
- Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, 63, 277-284.
- Godambe, V. P. (1985). The foundations of finite sample estimation in stochastic processes. *Biometrika*, 72, 419-428.
- Godambe, V. P. (1994). Linear Bayes and optimal estimation. preprint.

- Godambe, V. P. and Heyde, C. C. (1987), Quasi-likelihood and optimal estimation. *Int. Stat. Rev.*, 55, 231-244.
- Godambe, V. P. and Kale, (1991), Estimating functions: an overview, Estimating Functions. Ed. V. P. Godambe, Oxford University Press, New York, 2-30.
- Godambe, V. P. and Thompson, M (1974) Estimating equations in the presence of nuisance parameters. *Ann. Stat.*, 2 (3), 568-571.
- Godambe, V. P. and Thompson, M. E. (1989). An extension of quasi-likelihood estimation (with discussions). *J. Stat. Plan. Inf.*, 22, 137-172.
- Haines, L. M. (1995), A geometric approach to optimal design for one-parameter non-linear models. *J. R. Stat. Soc. B*, 57 (3), 575-598.
- Hoeffding, W. (1992), A class of statistics with asymptotically normal distribution. Breakthroughs in Statistics, Vol.1, 308-334.
- Huber, P. J. (1964), Robust estimation of a location parameter. *Ann. Math. Stat.*, 35, 73-101.
- Huber, P. (1980), Robust Statistics. John Wiley and Sons, New York.
- Heyde, C. C. (1989), Quasi-likelihood and optimality for estimating functions: some current unifying themes. *Bull. Inter. Stat. Inst.*, 1, 19-29.
- Karlin, S. and Studden, W. J. (1966), Optimal experimental designs. *Ann. Math. Stat.*, 37, 783-815.
- Kale, B. K. (1962). An extension of Cramer-Rao inequality for statistical estimation functions. *Skand. Aktur.*, 45, 60-89.
- Kiefer, J. (1959), Optimum experimental designs. *J. Roy. Stat. Soc. B*, 21, 272-319.
- Kiefer, J. (1974), General equivalence theory for optimum designs (approximate theory). *Ann. Stat.*, 2, 849-879.
- Kiefer, J. and Wolfowitz, J. (1959), Optimum designs in regression problems. *Ann. Math. Stat.*, 30, 271-294.
- Kiefer, J. and Wolfowitz, J. (1960), The equivalence of two extremum problems. *Can. J. Math.*, 14, 363-366.
- Kumon, M. and Amari, S. I. (1984), Estimation of structural parameters in the presence of a large number of nuisance parameters. *Biometrika*, 71 (3), 445-459.

- Laird, N. M. (1978), Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Stat. Assoc.*, 73, 805-811.
- Liang, K. Y. and Waclawiw, M. A. (1990), Extension of the Stein estimating procedure through the use of estimating functions. *Journal of American Statistical Association*, 85 (410), 435-440.
- Liang, K. Y. and Zeger, S. L. (1986), Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.
- Liang, K. Y. and Zeger, S. L. (1995), Inference based on estimating functions in the presence of nuisance parameters (with discussions). *Statistical Science*, 10, 158-199.
- Liang, K. Y. Zeger, S. L. and Qaqish, B. (1992), Multivariate regression analysis for categorical data (with discussion). *J. R. Stat. Soc. B* 54, 3-40.
- Lindsay, B. G. (1981), Properties of the maximum likelihood estimator of a mixing distribution. *Statistical Distributions in Scientific Work*, edited by G. P. Patil, Vol.5. Reidel, Boston, 95-109.
- Lindsay, B. G. (1982), Conditional score functions: some optimality results. *Biometrika*, 69 503-512.
- Lindsay, B. G. (1983a), The geometry of mixture likelihoods: a general theory. *Ann. Stat.*, 11 (1), 86-94.
- Lindsay, B. G. (1983b), The geometry of mixture likelihoods, Part II: the exponential family. *Ann. Stat.*, 11 (3), 783-792.
- Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications*. Institute of Mathematical Statistics, Vol. 5.
- Lloyd, C. J. (1987), Optimality of marginal likelihood estimating equations. *Comm. Stat., Theory and Meth.*, 16, 1733-1741.
- McCullagh, P. (1983), Quasi-likelihood functions. *Ann. Statist.*, 11 (1), 59-67.
- McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*. 2nd Ed. Chapman and Hall, London.
- McGilchrist, C. A. (1994), Estimation in generalized mixed models. *J. R. Statist. Soc. B* 56 (1), 61-69.
- McLeish, D. L. and Small, C. G. (1992), A projected likelihood function for semiparametric models. *Biometrika*, 79 (1), 93-102.

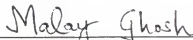
- Morris, C. L. (1983), Parametric empirical Bayes inference: theory and applications. *Journal of American Statistical Association*, 78 (381), 47-55.
- Murphy, S. and Li, B. (1995), Projected partial likelihood and its application to longitudinal data. *Biometrika*, 82, 399-406.
- Nelder, J. A. and Wedderburn, R. W. M. (1971), Generalized linear models. Breakthroughs in Statistics, Vol. 2, Springer-Verlag, New York, 547-563.
- Pazman, A. (1986), Foundations of Optimum Experimental Design. D. Reidel Publishing Company, Boston.
- Raghunathan, T. E. (1993), A quasi-empirical Bayes method for small area estimation. *Journal of American Statistical Association*, 88 (424), 1444-1448.
- Schall, R. (1991), Estimation in generalized linear models with random effects. *Biometrika*, 78 (4), 719-727.
- Serfling, R. J. (1980), Approximation Theorems of Mathematical Statistics. John Wiley and Sons Inc., New York.
- Shaked, M. (1980), On mixtures from exponential families. *J. R. Stat. Soc. B*, 42 (2), 192-198.
- Silvey, S. D. (1980), Optimal Design, Chapman and Hall, London.
- Small, C. G. and McLeish, D. L. (1994), Hilbert Space Methods in Probability and Statistical Inference, John Wiley and Sons, Inc., New York.
- Small, C. G. and McLeish, D. L. (1988), Generalization of ancillarity, completeness and sufficiency in an inference function space. *Ann. Stat.*, 16, 534-551.
- Small, C. G. and McLeish, D. L. (1989), Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, 76, 693-703.
- Studden, W. J. (1971), Elfving's theorem and optimal designs for quadratic loss. *Ann. Math. Stat.*, 42 (5), 1613-1621.
- Waclawiw, M. A. and Liang, K. Y. (1994), Empirical Bayes estimation and inference for the random effects models with binary response. *Statistics in Medicine*, 13, 541-551.
- Waclawiw, M. A. and Liang, K. Y. (1993), Prediction of random effects in the generalized linear model. *Journal of American Statistical Association*, 88 (421), 171-178.
- Wedderburn, R. W. M. (1974), Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika*, 63, 27 - 32.

- Whittle, P. (1973), Some general points in the theory of optimal experimental design. , *J. Roy. Stat. Soc., B*, 35, 123-130.
- Zeger, S. L. and Liang, K. Y. (1986), Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121-130.
- Zeger, S. L. and Liang, K. Y. (1992), An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11, 1825-1839.
- Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988), Models for longitudinal data: a generalized estimating equation approach. *Biometrics* 44, 1049-1060.

BIOGRAPHICAL SKETCH

Schultz Chan got his first Ph.D. in mathematical physics from the University of Iowa in August 1992, and came to the University of Florida as a postdoctoral research associate. After completing his postdoctoral appointment in August 1994, and realizing that he wanted to change careers, Schultz Chan came to the Statistics Department to study applied statistics. He is expecting to get his second Ph.D. this August and is ready to face the real world.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



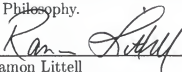
Malay Ghosh, Chairman
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Richard Scheaffer
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



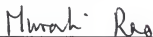
Ramon Littell
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



James Booth
Associate Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.



Murali Rao
Professor of Mathematics

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 1996

Dean, Graduate School